USENIX Association

Proceedings of the Nineteenth Symposium on Usable Privacy and Security (SOUPS 2023)

August 7–8, 2023
Anaheim, CA, USA

© 2023 by The USENIX Association

All Rights Reserved

This volume is published as a collective work. Rights to individual papers remain with the author or the author's employer. Permission is granted for the noncommercial reproduction of the complete work for educational or research purposes. Permission is granted to print, primarily for one person's exclusive use, a single copy of these Proceedings. USENIX acknowledges all trademarks herein.

ISBN 978-1-939133-36-6

Symposium Organizers

General Co-Chairs
Patrick Gage Kelley, *Google*
Apu Kapadia, *Indiana University Bloomington*

Technical Papers Co-Chairs
Katharina Krombholz, *CISPA Helmholtz Center for Information Security*
Rick Wash, *Michigan State University*

Technical Papers Committee
Ruba Abu-Salma, *King's College London*
Taslima Akter, *University of California, Irvine*
Florian Alt, *Bundeswehr University Munich*
Nalin Asanka Gamagedara Arachchilage, *The University of Auckland*
Hala Assal, *Carleton University*
Adam J. Aviv, *The George Washington University*
Rebecca Balebako, *Google*
Alexandru Bardas, *University of Kansas*
Eleanor Birrell, *Pomona College*
Jean Camp, *Indiana University*
Camille Cobb, *University of Illinois Urbana–Champaign*
Lynne Coventry, *Northumbria University*
Sauvik Das, *Carnegie Mellon University*
Pardis Emami-Naeini, *Duke University*
Cori Faklaris, *University of North Carolina at Charlotte*
Carrie Gates, *Bank of America*
Maximilian Golla, *Max Planck Institute for Security and Privacy*
Julie Haney, *National Institute of Standards and Technology (NIST)*
Rakibul Hasan, *Arizona State University*
Cormac Herley, *Microsoft Research*
Jun Ho Huh, *Samsung Research*
Patrick Gage Kelley, *Google*
Hyoungshick Kim, *Sungkyunkwan University*
Bart Knijnenburg, *Clemson University*
Janne Lindqvist, *Aalto University*
Heather Lipford, *University of North Carolina at Charlotte*
Sana Maqsood, *York University*
Abigail Marsh, *Macalester College*
Peter Mayer, *University of Southern Denmark*
Susan E. McGregor, *The Data Science Institute at Columbia University*
Mainack Mondal, *Indian Institute of Technology Kharagpur*
Maryam Mustafa, *Lahore Institute of Management Sciences*
Alena Naiakshina, *Ruhr-Universität Bochum*
James Nicholson, *Northumbria University*
Simon Parkin, *Delft University of Technology*
Emilee Rader, *Michigan State University*
Irwin Reyes, *Two Six Labs*
Franzi Roesner, *University of Washington*
Scott Ruoti, *The University of Tennessee*
Florian Schaub, *University of Michigan*
Kent Seamons, *Brigham Young University*
Matthew Smith, *University of Bonn*
Jessica Staddon, *JPMorgan Chase*
Elizabeth Stobert, *Carleton University*
Jose Such, *King's College London and VRAIN-UPV*
Blase Ur, *The University of Chicago*
Josephine Wolff, *The Fletcher School at Tufts University*
Heng Xu, *American University*
Yaxing Yao, *Virginia Tech*
Daniel Zappala, *Brigham Young University*
Leah Zhang-Kennedy, *University of Waterloo*
Mary Ellen Zurko, *MIT Lincoln Laboratory*

Invited Talks Chair
Joe Calandrino, *US Federal Trade Commission*

Lightning Talks and Demos Co-Chairs
Taslima Akter, *University of California, Irvine*
Kopo Marvin Ramokapane, *University of Bristol*

Lightning Talks and Demos Junior Co-Chair
Nikita Samarin, *University of California, Berkeley*

Karat Award Chair
Heather Lipford, *University of North Carolina at Charlotte*

Posters Co-Chairs
Hala Assal, *Carleton University*
Joshua Reynolds, *New Mexico State University*

Posters Junior Co-Chair
Kentrell Owens, *University of Washington*

Tutorials and Workshops Co-Chairs
Daniel Votipka, *Tufts University*
Yaxing Yao, *Virginia Tech*

Tutorials and Workshops Junior Co-Chair
Kelsey Fulton, *University of Maryland, College Park*

Mentoring Co-Chairs
Sana Maqsood, *York University*
Scott Ruoti, *The University of Tennessee*

Mentoring Junior Co-Chairs
Sabid Bin Habib, *Indiana University*
Lea Gröber, *CISPA Helmholtz Center for Information Security*

Publicity Co-Chairs
Martin Degeling, *Ruhr-Universität Bochum*
Yixin Zou, *Max Planck Institute for Security and Privacy*

Local Publicity Chair
Yaxing Yao, *Virginia Tech*

Publicity Junior Co-Chair
Jane Im, *University of Michigan*

Email List Chair
Lorrie Cranor, *Carnegie Mellon University*

Accessibility Chair
Liz Markel, *USENIX Association*

USENIX Liaison
Casey Henderson, *USENIX Association*

External Reviewers

Aniqa Alam
Mir Masood Ali
David Balash
Benjamin Berens
Arkaprabha Bhattacharya
Amel Bourdoucen
Yee-Yin Choong
Verena Distler
Matt Dixon
Marco Gutfleisch
Franziska Herbert
Sungsu Kwag
Leona Lassak
Henrik Lassila
Kevin Lee
Philipp Markert
Francesca Mosca
Collins Munyendo
Andy Regenscheid

Message from the SOUPS 2023 Program Co-Chairs

Welcome to SOUPS 2023!

With the conference in its 19th year, our SOUPS community has collectively ensured an excellent and exciting conference program. With 33 papers accepted out of 147 submissions (22% acceptance rate), the technical program covers a wide range of topics within usable privacy and security. The conference also includes workshops, posters, lightning talks, mentorship activities, and a keynote.

In 2016, SOUPS became an independent conference body. For the last six years, we have partnered with USENIX for hosting and administrative support, a move that has enabled continued growth for the conference. We thank all the members of the USENIX staff for their work in organizing SOUPS and supporting our community. Their team has been fantastic at making the process seamless.

In 2018, we co-located with the USENIX Security Symposium for the first time, and we have continued that co-location for 2023. Co-locating the two conferences allows for interactions and shared ideas between SOUPS and USENIX Security attendees. We have found this beneficial for both conferences and look forward to the opportunity again this year. After years of disruption from the pandemic, we are back to an all in-person format this year. We hope returning to this format will help you find SOUPS 2023 engaging and meaningful.

SOUPS relies on a range of volunteers for all of its activities. Steering Committee members provide oversight and guidance and are elected for three-year terms. Organizing Committee members help determine the conference content for a particular year, often serving two-year terms to facilitate the transition of knowledge. Technical Papers Committee members are chosen by the Technical Papers Co-Chairs each year. SOUPS is a product of the hard work by many people, starting with researchers who decide to submit their work to SOUPS, and including all of the SOUPS Organizers, the SOUPS Steering Committee, the technical paper reviewers, the workshop organizers, the poster jury, and the USENIX staff. We are grateful and thank each and every one of you for your contributions to SOUPS 2023.

Apu Kapadia has served as General Chair of SOUPS and Chair of the Steering Committee for 2023. Patrick Gage Kelley was appointed as Vice Chair in 2023. If you are interested in helping with SOUPS 2024 in any way, please contact Patrick.

SOUPS would not be possible without the generous support of our sponsors – thank you. Please visit our website to view the recipients of the SOUPS 2023 awards. Congratulations to all recipients for their outstanding work.

Patrick Gage Kelley, *Google, General Co-Chair*
Apu Kapadia, *Indiana University, General Co-Chair*
Katharina Krombholz, *CISPA Helmholtz Center for Information Security, Technical Papers Co-Chair*
Rick Wash, *Michigan State University, Technical Papers Co-Chair*

Nineteenth Symposium on Usable Privacy and Security (SOUPS 2023)

August 7–8, 2023
Anaheim, CA, USA

Monday, August 7

Cybercrimes and Misinformation

An Investigation of Teenager Experiences in Social Virtual Reality from Teenagers', Parents', and Bystanders' Perspectives 1
Elmira Deldari, *University of Maryland, Baltimore County;* Diana Freed, *Cornell Tech;* Julio Poveda, *University of Maryland;* Yaxing Yao, *University of Maryland, Baltimore County*

Fight Fire with Fire: Hacktivists' Take on Social Media Misinformation 19
Filipo Sharevski and Benjamin Kessell, *DePaul University*

"Stalking is immoral but not illegal": Understanding Security, Cyber Crimes and Threats in Pakistan 37
Afaq Ashraf and - Taha, *Lahore University of Management Sciences;* Nida ul Habib Bajwa and Cornelius J. König, *Universität des Saarlandes;* Mobin Javed and Maryam Mustafa, *Lahore University of Management Sciences*

Checking, nudging or scoring? Evaluating e-mail user security tools 57
Sarah Y. Zheng and Ingolf Becker, *UCL*

Understanding the Viability of Gmail's Origin Indicator for Identifying the Sender 77
Enze Liu, Lu Sun, and Alex Bellon, *UC San Diego;* Grant Ho, *University of Chicago;* Geoffrey M. Voelker, Stefan Savage, and Imani N. S. Munyaka, *UC San Diego*

Security and Privacy in Organizations

'Give Me Structure': Synthesis and Evaluation of a (Network) Threat Analysis Process Supporting Tier 1 Investigations in a Security Operation Center 97
Leon Kersten, Tom Mulders, Emmanuele Zambon, Chris Snijders, and Luca Allodi, *Eindhoven University of Technology*

Exploring the Security Culture of Operational Technology (OT) Organisations: The Role of External Consultancy in Overcoming Organisational Barriers 113
Stefanos Evripidou, *University College London;* Uchenna D Ani, *University of Keele;* Stephen Hailes and Jeremy D McK. Watson, *University College London*

Lacking the Tools and Support to Fix Friction: Results from an Interview Study with Security Managers 131
Jonas Hielscher, Markus Schöps, Uta Menges, Marco Gutfleisch, Mirko Helbling, and M. Angela Sasse, *Ruhr University Bochum*

What can central bank digital currency designers learn from asking potential users? 151
Svetlana Abramova and Rainer Böhme, *Universität Innsbruck;* Helmut Elsinger, Helmut Stix, and Martin Summer, *Oesterreichische Nationalbank*

"Would You Give the Same Priority to the Bank and a Game? I Do Not!" Exploring Credential Management Strategies and Obstacles during Password Manager Setup 171
Sabrina Klivan, *CISPA Helmholtz Center for Information Security;* Sandra Höltervennhoff and Nicolas Huaman, *Leibniz University Hannover;* Yasemin Acar, *George Washington University and Paderborn University;* Sascha Fahl, *CISPA Helmholtz Center for Information Security and Leibniz University Hannover*

Evolution of Password Expiry in Companies: Measuring the Adoption of Recommendations by the German Federal Office for Information Security 191
Eva Gerlitz, *Fraunhofer FKIE;* Maximilian Häring, *University of Bonn;* Matthew Smith, *University of Bonn, Fraunhofer FKIE;* Christian Tiefenau, *University of Bonn*

Authentication

Dissecting Nudges in Password Managers: Simple Defaults are Powerful 211
Samira Zibaei, Amirali Salehi-Abari, and Julie Thorpe, *Ontario Tech University*

Adventures in Recovery Land: Testing the Account Recovery of Popular Websites When the Second Factor is Lost 227
Eva Gerlitz, *Fraunhofer FKIE;* Maximilian Häring and Charlotte Theresa Mädler, *University of Bonn;* Matthew Smith, *University of Bonn, Fraunhofer FKIE;* Christian Tiefenau, *University of Bonn*

Tangible 2FA – An In-the-Wild Investigation of User-Defined Tangibles for Two-Factor Authentication 245
Mark Turner, *University of Glasgow;* Martin Schmitz, *Saarland University Saarbrücken;* Morgan Masichi Bierey and Mohamed Khamis, *University of Glasgow;* Karola Marky, *University of Glasgow and Ruhr-University Bochum*

Prospects for Improving Password Selection 263
Joram Amador, Yiran Ma, Summer Hasama, Eshaan Lumba, Gloria Lee, and Eleanor Birrell, *Pomona College*

Tuesday, August 8

Beyond End Users/Developers and Experts

Who Comes Up with this Stuff? Interviewing Authors to Understand How They Produce Security Advice 283
Lorenzo Neil, *North Carolina State University;* Harshini Sri Ramulu, *George Washington University;* Yasemin Acar, *Paderborn University & George Washington University;* Bradley Reaves, *North Carolina State University*

Towards Usable Security Analysis Tools for Trigger-Action Programming 301
McKenna McCall and Eric Zeng, *Carnegie Mellon University;* Faysal Hossain Shezan, *University of Virginia;* Mitchell Yang and Lujo Bauer, *Carnegie Mellon University;* Abhishek Bichhawat, *IIT Gandhinagar;* Camille Cobb, *University of Illinois Urbana-Champaign;* Limin Jia, *Carnegie Mellon University;* Yuan Tian, *University of California, Los Angeles*

On the Recruitment of Company Developers for Security Studies: Results from a Qualitative Interview Study 321
Raphael Serafini, Marco Gutfleisch, Stefan Albert Horstmann, and Alena Naiakshina, *Ruhr University Bochum*

SoK: I Have the (Developer) Power! Sample Size Estimation for Fisher's Exact, Chi-Squared, McNemar's, Wilcoxon Rank-Sum, Wilcoxon Signed-Rank and t-tests in Developer-Centered Usable Security 341
Anna-Marie Ortloff and Christian Tiefenau, *University of Bonn;* Matthew Smith, *University of Bonn, Fraunhofer FKIE*

Accessibility and Allies

GuardLens: Supporting Safer Online Browsing for People with Visual Impairments 361
Smirity Kaushik, Natã M. Barbosa, Yaman Yu, Tanusree Sharma, Zachary Kilhoffer, and JooYoung Seo, *University of Illinois at Urbana-Champaign;* Sauvik Das, *Carnegie Mellon University;* Yang Wang, *University of Illinois at Urbana-Champaign*

Iterative Design of An Accessible Crypto Wallet for Blind Users 381
Zhixuan Zhou, Tanusree Sharma, and Luke Emano, *University of Illinois at Urbana-Champaign;* Sauvik Das, *Carnegie Mellon University;* Yang Wang, *University of Illinois at Urbana-Champaign*

Youth understandings of online privacy and security: A dyadic study of children and their parents 399
Olivia Williams, *University of Maryland;* Yee-Yin Choong and Kerrianne Buchanan, *National Institute of Standards and Technology*

ImageAlly: A Human-AI Hybrid Approach to Support Blind People in Detecting and Redacting Private Image Content 417
Zhuohao (Jerry) Zhang, *University of Washington, Seattle;* Smirity Kaushik and JooYoung Seo, *University of Illinois at Urbana-Champaign;* Haolin Yuan, *Johns Hopkins University;* Sauvik Das, *Carnegie Mellon University;* Leah Findlater, *University of Washington, Seattle;* Danna Gurari, *University of Colorado Boulder;* Abigale Stangl, *University of Washington, Seattle;* Yang Wang, *University of Illinois at Urbana-Champaign*

Evaluating the Impact of Community Oversight for Managing Mobile Privacy and Security 437
Mamtaj Akter, *Vanderbilt University;* Madiha Tabassum and Nazmus Sakib Miazi, *Northeastern University;* Leena Alghamdi, *University of Central Florida;* Jess Kropczynski, *University of Cincinnati;* Pamela J. Wisniewski, *Vanderbilt University;* Heather Lipford, *University of North Carolina, Charlotte*

Beliefs and Behavior

Data Privacy and Pluralistic Ignorance 457
Emilee Rader, *Michigan State University*

Distrust of big tech and a desire for privacy: Understanding the motivations of people who have voluntarily adopted secure email 473
Warda Usman, Jackie Hu, McKynlee Wilson, and Daniel Zappala, *Brigham Young University*

"Is Reporting Worth the Sacrifice of Revealing What I've Sent?": Privacy Considerations When Reporting on End-to-End Encrypted Platforms 491
Leijie Wang and Ruotong Wang, *University of Washington;* Sterling Williams-Ceci, *Cornell University;* Sanketh Menda, *Cornell Tech;* Amy X. Zhang, *University of Washington*

Evaluating User Behavior in Smartphone Security: A Psychometric Perspective 509
Hsiao-Ying Huang, *University of Illinois at Urbana Champaign;* Soteris Demetriou, *Imperial College London;* Muhammad Hassan, *University of Illinois at Urbana Champaign;* Güliz Seray Tuncay, *Google;* Carl A. Gunter and Masooda Bashir, *University of Illinois at Urbana Champaign*

Privacy Mental Models of Electronic Health Records: A German Case Study 525
Rebecca Panskus, *Ruhr-University Bochum;* Max Ninow, *Leibniz University Hannover;* Sascha Fahl, *CISPA Helmholtz Center for Information Security;* Karola Marky, *Ruhr-University Bochum and Leibniz University Hannover*

Future Internet/Smart Home, the Metaverse, and AI

"Nobody's Happy": Design Insights from Privacy-Conscious Smart Home Power Users on Enhancing Data Transparency, Visibility, and Control 543
Sunyup Park and Anna Lenhart, *University of Maryland, College Park;* Michael Zimmer, *Marquette University;* Jessica Vitak, *University of Maryland, College Park*

Exploring the Usability, Security, and Privacy of Smart Locks from the Perspective of the End User 559
Hussein Hazazi and Mohamed Shehab, *University of North Carolina at Charlotte*

"There will be less privacy, of course": How and why people in 10 countries expect AI will affect privacy in the future 579
Patrick Gage Kelley, *Google;* Celestina Cornejo and Lisa Hayes, *Ipsos;* Ellie Shuo Jin, Aaron Sedley, Kurt Thomas, Yongwei Yang, and Allison Woodruff, *Google*

Investigating Security Indicators for Hyperlinking Within the Metaverse 605
Maximiliane Windl, *LMU Munich and Munich Center for Machine Learning* (*MCML*)*;* Anna Scheidle, *LMU Munich;* Ceenu George, *University of Augsburg and TU Berlin;* Sven Mayer, *LMU Munich and Munich Center for Machine Learning* (*MCML*)

An Investigation of Teenager Experiences in Social Virtual Reality from Teenagers', Parents', and Bystanders' Perspectives

Elmira Deldari*
University of Maryland, Baltimore County

Diana Freed*
Cornell University

Julio Poveda
University of Maryland

Yaxing Yao
University of Maryland, Baltimore County

Abstract

The recent rise of social virtual reality (VR) platforms has introduced new technology characteristics and user experiences, which may lead to new forms of online harassment, particularly among teenagers (aged 13-17). In this paper, we took a multi-stakeholder approach and investigate teenagers' experiences and safety threats in social VR from three perspectives (teenagers, parents, and bystanders) to cover complementary perspectives. Through an interview study with 24 participants (8 teenagers, 7 parents, and 9 bystanders), we found several safety threats that teenagers may face, such as virtual grooming, ability-based discrimination, unforeseeable threats in privacy rooms, etc. We highlight new forms of harassment in the social VR context, such as erotic role-play and abuse through phantom sense, as well as the discrepancies among teenagers, parents, and bystanders regarding their perceptions of such threats. We draw design implications to better support safer social VR environments for teenagers.

1 Introduction

Social virtual reality, also referred to as social VR, is a 3D virtual environment where users can interact with others through VR devices (e.g., VR headsets and controllers) [22, 41]. Social VR experiences are unique compared to those offered by other online spaces such as social media because of the fully immersive experience through voice, touching, and grabbing features using full-body or half-body tracking avatars [33]. Among all users, teenagers (between 13 and 17 years old) have become one of the largest user groups in social VR. Technology companies such as Meta are increasing their efforts to bring more teenagers to their social VR platforms as they represent the future of their user base [12].

Prior research has shown that teenagers face significant safety and privacy risks in social VR. For example, teenagers are exposed to violence, abuse, sexually explicit content, age-inappropriate content, voice trolling, and scaring, among others [35, 35, 48]. They are also exposed to traditional forms of bullying and name-calling, as well as unique forms of harassment that are specific to social VR, such as stalking individuals across rooms or worlds [36].

Despite the risks noted in prior literature, our understanding of teenagers' experiences with social VR and how to protect their safety, security, and privacy is still not comprehensive. We add to the literature by filling two significant gaps. First, prior research has been focused on a single perspective in social VR (e.g., users, teenagers, etc.). However, as a complex social environment, a typical social VR scene often involves multiple stakeholders, such as teenagers themselves and other adult users. This multi-stakeholder perspective has not yet been addressed. These stakeholders co-exist and may interact with each other in social VR. They may have different, even conflicting perspectives on their social VR experiences. Such perspectives may also have an impact on how they behave themselves and respond to other risks and threats. Second, unlike adult users who can purchase VR devices by themselves, most teenagers receive VR devices as a gift from their parents. A clearer understanding of whether the parents are aware of the potential threats and risks their children may encounter when using VR devices is much needed.

In this project, we take a multi-stakeholder approach to study teenagers' experiences in social VR from the perspectives of three distinct stakeholder groups: teenagers, bystanders, and parents. *Teenagers* include youth who are between 13 and 17 years old. *Bystanders* encompasses users in social VR who are not teenagers. *Parents* refers to parents of teenagers who are social VR users. We aim to study the following three research questions:

* Elmira Deldari and Diana Freed contributed equally to this research.

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.

USENIX Symposium on Usable Privacy and Security (SOUPS) 2023.
August 6–8, 2023, Anaheim, CA, United States.

- RQ1: What threats are teenagers exposed to in social VR from the perspectives of teenagers, bystanders, and parents?
- RQ2: What are the similar perspectives and tensions among teenagers, bystanders, and parents regarding social VR threats?
- RQ3: What features do teenagers, bystanders, and parents desire to combat safety threats in social VR?

To answer these research questions, we conducted an interview study with 8 teenagers, 9 bystanders, and 7 parents. The interviews focused on participants' experiences in social VR, their perceptions and perspectives of the safety threats, and their mitigation strategies when facing these threats. Our analysis shows that some activities, such as Erotic Role Play (ERP, a type of role-playing activity that includes users decorating their avatars with components that have sexual orientation), are present among teenagers, yet many teenagers seem to have normalized such activities, and did not consider them as threats. On the other hand, most bystander participants evaluate the activities in social VR using the norms from our physical world and identified many types of risks that may jeopardize teenagers' mental and physical health. Parents generally showed a limited understanding of the threats that their teenagers may face in social VR, with many being aware of only a few potential risks. Our results highlight the discrepancies among the perspectives of three stakeholders, which may lead to conflicting social norms in social VR and possibly more significant risks for teenagers.

Our paper makes two contributions. First, we explored teenagers' experiences and safety threats from the perspective of teenagers, bystanders, and parents. This multi-stakeholder approach allows us to comprehensively examine our research questions with complementary opinions and experiences. To the best of our knowledge, this is the first study to conduct interviews with three distinct groups with a particular focus on their interactions with others and the identification of potential threats in social VR. Second, this study provides insights and design implications that aim to create safer and more fulfilling social VR spaces for teenagers. By drawing from the perspectives of parents, bystanders, and teenagers themselves, these implications can inform the design of future social VR platforms and other online social spaces.

2 Related Work

2.1 Social VR: Benefits and Drawbacks

In social VR, users can create avatars that represent them in virtual spaces, then interact with others using their body gestures through full-body tracking (i.e., the body movement of a user's avatar corresponds to the body movement of the user in real-time) [7, 9, 57, 61]. This real-time embodiment allows users not only to customize their avatars but also pilot them with real-time gestures and motions [24]. In addition, social VR allows users to connect with each other and gather with friends from anywhere around the world and share experiences and activities that would never be possible in person [38,42]. For example, they can watch movies in Bigscreen or play and/or create games in Rec Room. Another platform called AltspaceVR, which shut down in March 2023, offered varied activities such as interacting with people, attending events, etc. [3].

On the other hand, previous studies have highlighted that users of social VR platforms have experienced unpleasant experiences or have seen inappropriate behavior in virtual spaces. For instance, Blackwell et al. conducted an interview study with bystanders and reported that embodiment and presence in VR spaces make harassment feel more intense, and some features such as synchronous voice chat or avatar movements could trigger the risk of potential harassment in social VR [8]. Also, the results of Shriram and Schwartz's quantitative survey indicate that harassment was occasional in social VR platforms and that those in female avatars reported experiencing it more [55]. Also, scholars have studied marginalized users and how verbal and non-verbal communication could lead to potential risks of online harassment [37].

Moreover, prior work has suggested that, among all users of social VR, children and teenagers are the most vulnerable [8]. This is due to the fact that interaction dynamics between adults and children in social VR introduce barriers, tensions, and frustrations due to the co-existence of mixed ages in this social space [35, 36]. Some adults have expressed concerns for younger users in social VR because of the prominent harassment risks [35]. Researchers have also observed incidents in which young people were exposed to inappropriate content such as sex, alcohol, and virtual sexual assault [34].

2.2 Technology-Facilitated Harassment

Among all issues teenagers may face in social VR, harassment is the one rising the quickest [14,30,49]. Abusive sexual behavior could have a profound impact on young people's mental and physical health (e.g., anxiety, distress) as well as on the development of their sexuality and social functioning, both in the short term and long term [10]. With the development of digital technologies, harassment and sexual abuse have also raised significant legal issues such as viewing or uploading indecent images of children or teenagers on the Internet or consuming of other child sexual abuse materials (e.g., text, images, child pornography, etc.) [27, 30, 63], cyber-bullying [20, 56, 60], and cyber-grooming [19, 32, 47]. Moreover, technologies may make it easier to initiate, escalate, and maintain abuse in various contexts [29, 45, 54], such as mobile devices [39], social media [45, 50], gaming [11, 29], etc.

Numerous efforts have been made to combat online harass-

ment in order to promote a safe environment for young users. For instance, some technology companies have designed and implemented various mechanisms to detect, prevent, and report sexual harassment [6, 26, 44, 58]. Research has also highlighted the opportunity to use automated computational approaches for risk detection to support children's online safety based on machine learning models [1, 2, 4, 15, 25, 46, 53]. Additionally, educational materials primarily targeted at parents have been developed to keep them informed about how their teens can stay safe when using social VR [43,52]. Researchers have also been studying other ways to encourage teens to take action when experiencing harassment, like seeking peer support [31].

In this study, we build on prior work and focus on understanding teenagers' experiences and safety threats from the perspective of teenagers, bystanders, and parents. These complementary perspectives uncover nuances around teenagers' threats and point at opportunities for designing safety features and ensuring a safer and healthier virtual environment.

3 Methodology

To answer our research questions, we conducted a semi-structured interview study with teenagers, bystanders, and parents. We detail the study methodology in the following sections. This study is approved by our university's IRB.

3.1 Participant Recruitment

We focused on recruiting three groups of users: teenagers (ages 13 - 17) who have experienced social VR, parents whose teenagers have used social VR, and bystanders (ages 18+) who actively engage on social VR platforms. In the context of this study, we use "bystanders" to denote individuals who are neither teenagers nor parents but may have witnessed other teenagers' interactions with others in social VR (similar to [16]). We do not consider bystanders in the physical world who may stand next to users who use VR devices . In total, we recruited 24 participants, including 8 teenagers (T), 7 parents (P), and 9 bystanders (B). Table 1 includes participant demographic information and their social VR experience. Overall, our participants represent diverse backgrounds in terms of their age, occupation, and location.

We posted our recruitment flyer on popular online forums (e.g., Reddit subforums such as r/VRchat, r/RecRoom, and r/Oculus), online communities (e.g., Discord), and interest groups on social media sites (e.g., Twitter and Facebook). Before posting it to these sites, we sent our flyer and IRB approval letter to the corresponding platform/group moderators for their review. We only posted the flyer after obtaining the moderator's approval.

Candidates who were older than 13 years and were interested in our study were invited to fill in a screening survey through the link provided in our flyer. In the screening survey, we asked about their social VR experiences, the VR headset devices they have used, their frequency of using social VR, their ages, whether they have children, and if so, their children's ages.

Candidates with stable access to a VR headset and social VR experience were eligible to participate as either teenagers or bystanders. Teenagers were those who were aged between 13 and 17. Bystanders were general users in social VR who are 18+ (we used "bystanders" rather than "users" as we were interested in their experiences as bystanders of teenagers' activities). Candidates who had both access to a VR headset and teenagers in their household who used social VR were assigned to the "Parents" group. Although not required, all parents in our study had at least one year of social VR experience.

We did not limit our recruitment to certain geographic areas, as most social VR applications provide public places that can be accessed by users from any region in the world. We required participants to be able to communicate in English.

3.2 Interview Protocol

To accommodate the three participant groups, we framed the same interview protocol differently to account for the three different perspectives. Below, we describe the interview flow using the teenager version as an example.

The interview protocol consists of three parts. The first part focuses on the participants' background information (age, gender, etc.), their general VR experience, and their perceptions on social VR including their perceived benefits and concerns. The second part focuses on participants' behaviors and activities in social VR. We ask about their interactions with other users in social VR and how they approached/were approached by them. We then ask participants why they interact with other users and what their criteria are when they chose friends in social VR. In the next section, we focused on the risks and harms of social VR. We ask participants to share any negative experiences they encountered. Based on the participant's responses, we would either follow up with questions asking for more details or, if they did not have any negative experiences or could not think of any, we would ask whether they have encountered or witnessed any negative incidents, their opinions, and their reaction or strategies to navigate through those experiences. In the last section, we ask them whether they would like to see any features on existing social VR platforms.

3.3 Data Collection and Analysis

We conducted remote interviews via Zoom. The average interview length was 60 minutes and participants who completed the study received monetary compensation of USD $20 (or the equivalent value in their local currency). All interviews were

Group	ID	Gender	Age	Occupation	Location	Num. Kids	Usage Experience	Used Social VR Platforms
Teenager	T1	Female	17	Student	USA	0	2 years	VRChat, Rec Room Horizon Worlds
	T2	Male	14	Student	USA	0	2 years	VRChat, Rec Room
	T3	Male	17	Student	USA	0	1 year	VRChat, Rec Room
	T4	Male	14	Student	USA	0	2 years	Rec Room
	T5	Male	15	Student	Lithuania	0	2 years	Rec Room, EchoVR
	T6	Male	13	Student	USA	0	1 year	Rec Room
	T7	Male	13	Student	USA	0	1.5 years	VRChat
	T8	Female	17	Student	Belgium	0	2 years	VRChat, Rec Room
Bystanders	B9	Non-binary	47	Full-time employee	USA	0	1 year	AltspaceVR
	B10	Female	20	Caretaker	USA	0	2.5 years	VRChat, Rec Room
	B11	Male	21	Student	USA	0	1 year	VRChat, Rec Room
	B12	Female	22	Student	USA	0	1.5 years	VRChat, Rec Room HorizonWorlds, ChilloutVR
	B13	Male	23	Music instructor	Canada	0	5 years	VRChat
	B14	Female	21	Dance teacher	Canada	0	3 years	VRChat, Rec Room ChilloutVR
	B15	Male	20	Student	Japan	0	3 years	VRChat, Rec Room, Horizon Worlds, ChilloutVR
	B16	Female	NA	ASL teacher	USA	0	3 years	VRChat
	B17	Male	23	IT engineer	Brazil	0	1.5 years	VRChat, ChilloutVR
Parents	P18	Female	29	Lab manager	USA	1	1 year	AltspaceVR
	P19	Female	37	Housewife	USA	8	1 year	Rec Room
	P20	Male	41	Teacher	USA	1	1 year	VRChat, Rec Room
	P21	Male	35	Software engineer	Hungary	2	5 years	VRChat, Rec Room
	P22	Male	45	Architecture	Germany	2	2 years	VRChat
	P23	Male	53	IT project manager	UK	2	2 years	AltspaceVR, Bigscreen
	P24	Male	35	Pharmacist/ASL teacher	USA	1	3 years	VRChat, AltspaceVR

Table 1: Participants' demographics and social VR experience

audio-recorded upon participant consent and were then transcribed using Zoom's live transcription feature. We stopped the interviews when we did not observe new findings across all participant groups. As our study specifically focused on gathering teenagers' experiences from various perspectives, we reached saturation with a relatively small number of participants.

Next, one researcher manually cleaned all transcriptions by correcting all mistakes generated. We then conducted a thematic analysis to identify repetitive patterns and themes in the interviews. Three researchers first selected one random transcription from our teenager participants as a sample. They closely read through the sample data several times to immerse themselves in the data, and then coded the sample independently at the sentence level using open coding. Upon completion, the three researchers discussed the coding results together and generated an initial codebook. They then repeated the same process on two additional samples, one from the bystander participants and the other from the parent participants. Through this process, the research team generated 3 separate codebooks, one for each participant group.

Following this initial coding, three researchers separately coded the remaining data using the agreed codebook. New codes that emerged from the data were added. In this process, the research team met frequently to discuss the coding results, and updated the codebook as needed. This process was done

iteratively until all data was coded and full agreement was reached on the data from all three participant groups. All researchers then discussed and identified the themes for each user group.

Since our coding process involved multiple iterations and discussions and reached a full agreement, intercoder reliability was not necessary [40]. Upon completing the thematic analysis, the research team further compared the themes across all three participant groups.

3.4 Ethical Considerations

Since our study involved teenager participants, we took extra caution to ensure research ethics throughout the project, as described in detail below.

First, we asked all teenagers to obtain a parent's written consent before they could participate in our study. When we identified a qualified teenager from the screening survey, we sent them an assent form to sign together with a consent form for their parent to sign. To ensure that their parent was aware of their child's participation, teenagers were permitted to participate in the interview study only if they returned both signed assent and consent forms.

Second, before an interview with a teenager started, we always asked for separate oral consent from their parent. This is to verify that the teenager participants had indeed obtained their parent's permission to participate in our study.

Third, similar to the work done by Cranor et al. [13], when a teenager and their parents all reached out to us, we deliberately selected either the teenager or one of the parents to participate in our study (i.e., we only selected one participant from each household, thus the teenager participant and parent participant were not in pairs). This intentional setup was to 1) respect the teenager's right to privacy, especially if they did not want to share their experiences/opinions with their parents; and 2) avoid potential embarrassment or conflicts among family members after participating in our study. 3) When the participants shared their experiences in social VR, especially those that were deemed to be sensitive (e.g., experiences related to harassment), we reassured them that their responses would be kept anonymous. We also instructed participants that they could skip any questions if they preferred and doing that did not influence their compensation.

3.5 Limitations

Our study has various limitations. For instance, we only interviewed 8 teenagers, 7 parents, and 9 bystanders who are English speakers. While we believe that our sample size is sufficient for our study, we recognize that there may be other types of safety incidents experienced by teenagers in social VR that are yet to be discovered. Additionally, we did not interview parents and children from the same family together to understand family dynamics. As mentioned above, we intentionally chose not to do so for ethical considerations.

4 Results

In this section, we present our findings on teenagers' social VR experiences. We focus on teenagers' experiences and potential safety threats from three perspectives: teenagers, bystanders, and parents. This section follows the four major themes we identified in our data analysis, including participants' general perceptions of social VR, teenagers' relationship-building practices in social VR, teenagers' safety threats, and desired features. Given the qualitative nature of our study, when reporting the results, we used the terms "a few", "some", "several", "many", and "nearly all" to convey the relative sense of frequency rather than using specific numbers, similar to prior work [18, 28, 62].

4.1 Participants' General Perceptions of Social VR

Our participants from the three user groups demonstrated a consistent perception of social VR. Nearly all participants used social VR apps as a leisure activity to socialize, play games, and have intimate relationships in an immersive environment. Rec Room, VRChat, and AltspaceVR remain the most popular platforms among our participants. They were particularly drawn by several unique features of social VR platforms, such as real-time interaction, facilitating multimodal communications (e.g., through voice, tone, body movement, facial expression, etc.), and the lifelike social environment. Additionally, many participants indicated that the full-body movement and the ability to support fluid non-verbal communication alongside verbal communication contribute to the unique experiences and made it more genuine to engage in various activities. These results echoed the findings from several prior work [21, 22, 36, 37], thus we only summarize them briefly. In the following sections, we focus on the nuances of this study and show teenagers' experiences from the perspectives of teenagers, bystanders, and parents.

4.2 Building and Maintaining Relationships in Social VR

Compared to traditional 2D social networks, social VR provides a unique yet complex social environment, making it more challenging for teenagers to navigate through it. One common and fundamental activity relates to relationship building in social VR. Many teenager participants discussed how they have built and maintained relationships with other users in social VR, while many bystander and parent participants provided their observations to further uncover teenagers' practices.

In particular, while half of the teenager participants were able to bring their real-life friends into social VR for fun and interactive activities, the rest of them sought connections with new people. As a result, these teenagers were constantly involved in frequent and spontaneous interactions with strangers (i.e., people they have never met in real life). In this section, we present teenagers' strategies to develop and maintain relationships as well as their strategies to protect their own safety.

4.2.1 Various Strategies to Make Friends

Being in a complex social environment in social VR, teenagers have developed their own strategies for building connections with strangers. When approached by other users, teenagers relied on several signals to decide whether to respond or not.

Appropriate avatar behaviors as a positive sign. With limited information available to judge other users' characteristics, their behaviors became the primary factor in determining whether one would be accepted as a friend in a virtual world. The majority of the teenager participants reported that they preferred to make friends with those who exhibit decent and appropriate behavior. For example, T6 (13, male) mentioned that he may look for individuals who appeared to be respectful, kind, helpful, and avoid engaging in inappropriate or offensive behavior:

"I talked to them if they helped me with something, but if they're rude, I normally try to stay away from them, and most of the time in Gorilla tag, there's this button where you can mute people so that you don't have to listen to them." T6 (13, male)

As T6's example highlights, interacting with others in social VR could be a complex and challenging experience. He developed strategies for interacting with others that prioritize his own comfort and safety. Furthermore, T6 took proactive measures to protect his own well-being such as muting rude people in social VR to create a safe environment for himself. In general, our teenager participants selected who they talk with and chose to engage with people who are respectful and not prone to use rude or offensive language.

Many bystanders and parents in our study agreed that teenagers' safety should be the top priority. Yet, as adult users, bystanders and parents often focused more on engaging in interesting conversations when they themselves were users.

Seeking peers from the same age group. Furthermore, nearly all teenagers preferred to interact with a certain age group in social VR. as most teenagers often felt a greater sense of safety and comfort in forming friendships with users of the same age due to their shared experiences, common interests, and mutual understanding that come with being at a similar developmental stage.

"I feel like it's just easier to talk to my age. Because they just usually play for fun portion and then the older group I feel like it's just harder to talk to. Because they're just not the same age, so they can't relate to the things I do." T5 (15, male)

This perspective was further confirmed by many parents and bystander participants. For example, several parents mentioned that it is safer for their kids to interact with their own age group and peers. For example, P19 (37, female) commented:

"I want my kids to kick it with their peers in virtual reality, keep them safe and happy, by encouraging our children to become friends with individuals who are their own age or who they already know, we can provide them with a greater sense of security and comfort in these virtual environments." P19 (37, female)

In this quote, she emphasized the importance of parental involvement in keeping children safe and happy in social VR and she suggested that parents encourage their children to form friendships with individuals who are of their own age. By doing so, children could establish clear boundaries for communication and minimize the potential risks associated with strangers interacting online. However, It should be noted, that judging a user's age through their avatar is very challenging, as in most cases, there is no reliable indicator of a user's age in their avatar. A user's voice can be a reference, although mistakes can still occur. We will further unpack this point in the discussion.

Migration to cross-platforms to extend friendship. As social VR remains a synchronous platform, maintaining relationships becomes more difficult if the other users were not online. Thus, among many teenager participants, it was very common to migrate their interaction from social VR to other platforms (e.g., Discord), as they believed Discord offers a more convenient way to communicate with friends and sustain their relationships outside of the virtual environment. Furthermore, Discord's features to allow users to hide their identity and personal information, as well as the option to block individuals who make them feel uncomfortable or unsafe, provided a sense of control and security that is highly valued by many teenagers. T7 (13, male) commented on his experience with Discord:

"I decided to get Discord because it was what my Rec Room friends were using, and I just got it. And then I was like, hey I like this. Now I spend a lot of time talking to my friends about this. I'll never give them my number or email. Because that's, like personal. But Discord, I feel like you can still hide your identity." T7 (13, male)

On the contrary, a few parents believed that using Discord may cause additional risks to teenagers' safety since they believed that teenagers tend to share their personal information more easily on Discord, which could potentially lead to further risks. P20 commented:

"I was worried about my kid using Discord. I heard about these predators on the internet that try to get kids to give them

their personal information. And I thought, what if my kid gets caught up in that." P20 (41, male)

It is important to highlight that many other parent participants were not aware of the extended communication through these external platforms. This discrepancy made it challenging to maintain teenagers' safety. While most teenagers preferred to use other platforms to continue engaging with the people they met in social VR and believed it would be safe to do so, there was a lack of attention to these platforms from the parent's perspective. We will further discuss this phenomenon in the discussion section.

4.2.2 Casual Activities to Enhance Relationships

Social VR offers a unique and immersive experience that makes many seemingly unlikely social interactions possible in a virtual world. Nearly all teenagers in our study discussed their experiences of many different activities, such as playing games, dancing, sleeping, etc. Among these activities, some teenagers believed that casual activities (e.g., watching movies, having virtual parties, etc.) were effective ways to enhance the relationship among different users.

One popular activity that has been witnessed or experienced by multiple bystanders and parents is virtual drinking. To engage in this activity, one would enter a virtual bar that simulated the experience of a real-life bar, allowing them to socialize and spend time with their friends in a simulated bar environment. Essentially, virtual drinking events inherently serve as a social gathering that facilitates connections among users. However, some of the bystander and parent participants have expressed concerns about the involvement of teenagers in these events, as these drinking events were open to all ages and may nudge teenagers to drink in real life. Even though they have not yet seen such incidents happening to their teenagers, their concerns still exist. For example, P24 (35, male) stated the appropriateness of the situation, especially in the context of teenagers potentially being exposed to adults getting drunk in social VR:

"I see a lot of adults in a lot of the drinking worlds, for example, like the party drinking worlds, a lot of people seem to have a really really hard problem with either alcoholism or addiction [...] I worry about kids, as well, you know, because kids are impressionable, and this game is filled with predators. There are plenty of people who will take advantage of kids while they are drunk, just in general." P24 (35, male)

Furthermore, some parents further commented that those who got drunk in social VR environments may engage in behaviors that would be dangerous or intolerable in the real world, such as harassment, which could be especially harmful to teenagers. They may engage in inappropriate behaviors that could harm or exploit children, such as sharing inappropriate or explicit content or asking for personal information.

4.2.3 Safety Measures

As some teenagers appeared to be aware of the risks of connecting with virtual strangers, they have developed and adopted some measures to ensure their safety.

Use alternative identifications. One safety measure that several teenagers reported employing was being cautious about sharing their personal information. For example, in T1's (17, female) example, her approach of not sharing her name with strangers was an effective way to protect her personal information and maintain distance from individuals she did not know:

"I feel like I can trust strangers to a certain level, but I'm not fully trusting. I'm not gonna tell my name. I'll normally just have my friends call me by my first initial when I'm online. That is a common thing." T1 (17, female)

From some parents' perspective, they were concerned that teenagers might not be able to properly manage the distance with strangers and would possibly reveal personal information, which may further lead to great risks. Some parents confirmed such risks when interacting with strangers. P21 (35, male) shared his daughter's experience when she interacted with a stranger (an adult) who tried to communicate with her. In this case, he referred to the stranger as a "predator":

"My daughter was in the VRChat and people asked her for her address and if she has Facebook or Instagram. I don't want to judge anything, but at that moment, I thought there may be a pedophile, preying on children. Like what grown men ask like a child for Instagram and addresses just for friendship?" P21 (35, male)

Using avatars for anonymity. Most of the social VR platforms provide a variety of avatar options, including humanoid avatar (e.g., AltspaceVR, VRChat, Bigscreen) or non-humanoid avatar like an animal, superhero, or historical figure, or customized avatars from third-party platforms (only supported in VRChat), etc. [61]. This is, for the most part, designed for users to represent themselves in social VR. Some teenager participants agreed that social VR avatars could facilitate friendships by creating a visual representation of users that can be interacted with, allowing for greater immersion, social presence, and connection between users. Additionally, avatars facilitated nonverbal communication, such as gestures and body language. This is particularly important for conveying emotions, which are an essential aspect of human communication and are often difficult to express through text-based interactions.

Interestingly, using avatars may also create a sense of safety for some teenagers. In our dataset, several teenagers mentioned that avatars could provide people with a degree of anonymity and allow them to express themselves freely without revealing their real identities. This sense of anonymity made them feel more comfortable and less self-conscious, enabling them to build relationships with others more easily.

As T1 (17, female) mentioned, using avatars made her feel safer:

"I feel safer because it's not really a high risk. You don't really know who I am, you don't know where I live. You don't know what it looks like, it just feels safer having those cool avatars to represent you!" T1 (17, female)

4.3 Teenagers' Safety Threats

Prior work has suggested various types of threats in social VR, such as sexualized language, hate speech, visible sexual gestures, and so on [8, 23]. We continue to explore the safety threats that teenagers may face. In particular, our multi-stakeholder approach allowed us to explore not only teenagers' experiences but also the observed incidents from bystanders' and parents' perspectives. As a result, some of the following threats were reported by teenagers directly while others were observed by either bystanders or parents.

4.3.1 Sexual Harassment Through Erotic Role-Playing

Erotic Role-Play, or ERP, is a type of role-playing activity performed mostly or exclusively for sexual behavior and intentions. To do this, users would customize their avatars and decorate their avatars with symbols or components that have a sexual connotation.

Our teenager participants did not report their own experiences with ERP. However, some bystanders and parents repeatedly reported examples of ERP based on their experiences and how teenagers were engaged in ERP-related activities in social VR. They raised concerns about teenagers' access to adult-only ERP chats and content, such as virtual sex, lap dancing, etc. These activities were designed only for adults and would need to be accessed through private links on external channels (e.g., on Discord). However, these external channels were not associated with social VR applications and thus, were not restricted by the policies on social VR apps. As a result, teenagers were able to access such content easily.

For example, B11 (21, male) shared that while ERP activities were not published in public rooms, teenagers could still access them through quick searches in Discord channels or similar platforms, after which they would then ask for an invitation. He shared the time when he learned a teenager who got involved in ERP from a report :

"I follow some reports. I think he was just a 15-year-old who reached out to someone who did ERP [through Discord], and he released his age to the person in the ERP but still went through it. They allowed him to have some sort of ERP, even knowing his age." B11 (21, male)

Another example further suggested an alarming fact that even though the harassment activities happened in social VR, they started from places other than social VR. The safety measures and policies in the social VR platforms, regardless of their effectiveness, did not cover these external spaces, which may cause invisible threats to teenagers. Another bystander commented on this point:

"I think that it's accessible because it's as easy as a click of a button. If a teenager found out that there's a community for ERP or lap dancing, they could join the discord and figure out how to get in or something." B12 (22, female)

Relatedly, to enable erotic role-playing (ERP), one would need to have customized avatars through third-party platforms/software (e.g., Blender, Unity), and then import their avatars to social VR platforms (e.g., VRChat). Users are not obligated to adhere to any specific rules regarding the appearance of their personalized avatar on third-party software unless they need to meet certain technical requirements (e.g., rigging, polycount, textures, materials, and model format). Thus they are free to use any design, such as sexual components, insulting language, etc. These avatars may be inappropriate for teenagers to be exposed to.

4.3.2 "Feel" Virtual Harassment Through Phantom Sense

Phantom sense is a phenomenon caused by immersion in a VR environment where a user's brain tricks their physical body into feeling touch sensations on their virtual body in virtual environments. This phenomenon arises from the mind's confusion between reality and the virtual world. For example, when a user gets close to a fire in VR, their body will feel the heat. Usually, users can trick their minds to believe it is real and gain the ability to actually "feel" things in VR. Generally, there are different types of phantom senses - touch, smell, warmth, pain, etc., and every user can feel them, but some are more susceptible than others. It should be noted that with proper training, a user can make their phantom sense stronger and start feeling things and objects inside virtual reality.

While phantom sense can be used to intensify the emotion and joy of social VR activities, it may be misused by some malicious users for their own advantage. In our study, some teenager and bystander participants reported their experience of being harassed through phantom sense. T8 (17, female) shared her example:

"I have phantom sense on my arm, forehead, and nose too. It's not good to have it though. I regret mentioning I had it. If people know about it, a lot of them will abuse me. It feels like someone is scratching me, it's itchy ... I took off my headset like it makes me feel uncomfortable when they get close." T8 (17, female)

Relatedly, a few bystander participants reported an alarming fact: they reported that some teenagers took advantage of phantom sense and harassed other users without knowing the real consequences of it. For example, B10 (20, female) observed that when a female user talked about her phantom sense, a few other teenage boys in Rec Room started to touch

her body. This particular incident becomes alarming since teenagers, without proper guidance and rules in social VR, may flip their role from victims to predators without realizing it. She explained:

"I know quite a few people whose phantom sense becomes second nature to them to feel the things that they see happening to them. And don't ever say that you have phantom sense, because teenagers will do things to you against your will. I've seen it happen so many times in Rec Room that someone's talking about her phantom sense and as soon as you hear that everyone flocks to that person trying to find out who has it. They start touching her boobs, they start trying to rub her down there. They try kissing her or touching her neck." B10 (20, female)

In B10's example, she highlighted that teenagers may not have the maturity to regulate their behaviors in social VR, which could cause a risk for others who have a phantom sense to feel hurt in the physical world.

4.3.3 Physical Aggression

Virtual physical aggression. Physical aggression is behavior causing or threatening physical harm toward others. It includes hitting, kicking, biting, using weapons, and breaking toys or other possessions [17]. In our study, some teenagers reported various cases in which they were involved in physical aggression. For example, T5 (15, male) explained his experience with a team-based game in social VR:

"It's a team-based game where four people versus the other four people and I'm on one team and I kill one of their teammates, and the teammate starts being toxic and stuff, and the whole team just targets me, and hits my avatar, only because I killed their teammate." T5 (15, male)

In this case, neither our participant nor the other players in the game were physically hurt. However, the experience that our participant went through was disturbing. Such incidents became even more concerning considering the interconnection between physical aggression and violent behaviors, as research has shown that exposure to violent VR content could lead to elevated levels of aggression [51], posing long-term impacts on teenagers' mental health.

Parents normalize physical aggression. Interestingly, some parent participants held a different opinion regarding such physical aggression. They seemed to have normalized physical aggression and considered it as a normal aspect of playing virtual games. For example, P22 (45, male) mentioned that such behavior should be accepted as part of the gaming experience:

"So far, the only thing they [my kids] told me is that their thought on somebody who destroyed their house in Minecraft. Stuff like that happens in gaming. So somebody beat them in a game all the time and they were angry, but that's normal. " P22 (45, male)

P22 later suggested that he was also aware of the potential negative impacts of aggression and was taking steps to address it by asking his kids to share their experiences with him. As researchers, we believe that more active actions are needed to stop aggression from happening, as exposure to aggression in video games can have negative effects on children's behavior and social development [5]. We will further unpack this point in the discussion section.

4.3.4 Virtual Grooming Using Avatars

Grooming is one particular type of threat that can be difficult to identify by teenagers, as they are typically the victims without realizing it. Grooming refers to the situation in which an adult manipulates or abuses children or teenagers through building relationships and trust [19,32,47]. In our study, some bystanders shared their observations which they considered as grooming. For example, B10 (20, female) shared an example in which she unsuccessfully tried to help a 6-year-old boy:

"[PlayerID] admitted that he was looking for younger girls to be friends with, and he was 35. He had this 6-year-old, eating out of his hand. He groomed her into thinking that he was her friend and that she could only trust him, and I tried to help her. I tried to tell her this guy is a predator but she didn't believe me, she was too far gone." B10 (20, female)

This was an alarming example. Existing social VR platforms generally have a suggested age limit for their users (e.g., the age requirement for VRChat is 13 years or older [59]). Yet, children younger than 13 still accounted for a large percentage of the user base. Thus, users with malicious intentions may easily take advantage of them through grooming. Furthermore, avatars hide the real identity of the people behind them, making it difficult to identify the adults and their intention. As a result, trust can be built through some innovative ways, such as using a child's favorite avatar. P18 (29, female) provided an example:

"I just feel in a virtual reality setting, kids are more susceptible to manipulation. You can make that avatar something similar to a character that the younger kids would love. I would expect that to happen in VR, and any form of that, I would consider abuse and manipulation." P18 (29, female)

4.3.5 Potential Threats in a Private Room

In social VR, users can create or join private rooms, which are invitation-only spaces. The purpose of these private rooms is to provide a more controlled environment for users to interact and engage in activities. To understand the dynamics in private rooms, it is necessary to talk to people who have experiences in these rooms. In our study, several teenagers, parents, and bystanders have been invited to join private rooms, and their experiences pointed to potential threats to teenagers' safety. For example, P23 (53, male) shared a case in which

teenagers were invited to watch adult content in a private room in Bigscreen (a social VR app that supports movie sharing) and faced unforeseeable risks:

"I've seen it [adult content] quite a few times on Bigscreen. Adults will ask a child to join them in a private room and send a link to it. Or they'll open a room, then make it private when you're in there [...] I've seen porn movies in open rooms in Bigscreen. They're supposed to be safe so children don't see them. But they're not." P23 (53, male)

When facing threats in private rooms, some teenagers were able to identify, then responded proactively to combat the threats. For instance, T5 (15, male) witnessed a situation in which an adult user tried to lure a teenager into a private room in Echo VR. He quickly recognized the threat and took immediate action by reporting this adult user:

"I was chatting with a guy in Echo VR and an older man teleported in and talked to another player who was a boy. When I got closer to them, the older guy went quiet, he was trying to take the young kid to the private lobby to keep talking to him after I confronted him, I reported him." T5 (15, male)

As suggested by these examples, in private rooms, teenagers' threats and possible mitigation strategies may not be obvious to users and remain ineffective. We will discuss the implications in the discussion section.

4.3.6 Ability-Based Discrimination

Occasionally, the social VR environment was lacking inclusiveness, as reported by our participants. A few teenagers reported that they have witnessed incidents in which other users discriminated against some teenagers with disabilities. They highlighted the possibility that those users may not recognize the challenges that teenagers with disabilities may experience in social VR, and their seemingly joking behaviors may lead to discrimination, which negatively impacted the experiences of teenagers with disabilities. For example, a teenager described an incident in which another teenager with a speech disorder was discriminated against by other users:

"I've seen a kid that had speech disorders or speaking disabilities, he did speak weirdly, like he did not spell some word properly and they would go up to him and would make fun of him and ask why he has it." T7 (13, male)

4.4 Desired Safety Features in Social VR

Similar to previous studies that have focused on marginalized users (e.g., members of the LGBTQ community or women) who have used nonverbal communication (e.g., specific gestures) to protect themselves from potential harassment, [37, 48], our study explored several safety practices commonly used by our participants, such as reporting to the platforms, banning/muting/blocking other users, making other users invisible (e.g., using Personal Space Bubble), and assessing other users' trustworthiness before interacting with them (e.g., through Trust Rank). In this section, we present some nuances regarding participants' desired safety features.

Age matching mechanism. Many teenagers and bystanders often preferred to interact with others from a similar age group, while some parents preferred their teenagers to do the same. Our participants believe that matching players in public games based on their age may have a very positive impact on the social VR community by reducing undesired safety risks and harassment. As a teenager illustrated:

"I would probably try to separate it, like having two dedicated games, one for children and then one for adults. I think that would help mitigate harassment or even make it even easier to track who's harassing who and maybe make the discipline." T6 (13, male)

This age matching system was one of the most favored safety features by the majority of participants, yet it is important to acknowledge that implementing such features could inadvertently provide opportunities for predators. For example, predators could potentially exploit the system by reporting to be a kid, gaining approval from the platform, and then accessing a kid-only environment.

Age verification. To facilitate the age-matching process, another relevant safety feature is age verification. Some participants were looking for a feature in social VR platforms to ensure that a predator could not fake their age to access children's rooms and vice versa. For example, P24 (35, male) suggested that the platforms should ask for photo ID to confirm the identity and the age of the users:

"I'm hoping for a way to verify your age. So like a passport or something to verify that you're actually over the age of 13, so that minors don't get targeted by older audiences. I feel like kids should play with other kids and then everybody should play with their own age group." P24 (35, male)

Sexual harassment history indicator. As mentioned above, a banned user may create a new account and continue using the service. One safety feature that could remediate this issue would be to have an indicator on users' avatars regarding their harassment history. For example, P23 (53, male) suggested that the avatar of a previously banned user may include an indicator (e.g., a badge) to show their prior harassment record in the platform as a warning to other users:

"I feel like those who have been banned for sexual assault, or sexual behavior in a public area, should have some kind of mark on them [their avatars]. Like a sexual predator predictor. I feel like that would definitely help the community and maybe even discourage sexual assault." P23 (53, male)

This feature was highlighted by a few participants as a potential strategy to identify predators. While it may initially appear to be a promising approach, it is crucial to recognize that its implementation could inadvertently raise new forms

of harassment within the platform. For instance, users may specifically target or launch attacks against individuals who display this indicator, resulting in unpredictable consequences.

Parental control and involvement. Some participants highlighted that a significant part of child safety lies with their parents. Thus, some participants recognized the importance of having parental controls, such as limiting children's playing time, limiting the number of social VR platforms used by their children, etc. Our participants also suggested that parents need to be more engaged in their children's activities and be aware of the people they socialize with in social VR.

5 Discussion

As we move into an increasingly digital world, the realm of virtual reality (VR) has become an important area of focus for Human-Computer Interaction (HCI) studies. The rise of social VR presents new and unique challenges, particularly regarding the potential risks associated with its use. While previous HCI studies have explored these risks, it has been noted that most of these studies have only focused on a single group of users, such as young adults or bystanders [8, 23, 55].

Our study endeavors to explore teenagers' social VR experiences from the perspectives of teenagers, bystanders, and parents, who are all essential stakeholders in social VR ecosystems. This multi-stakeholder approach takes advantage of the unique experiences and perspectives of each stakeholder and provides different yet complementary angles to understand teenagers' experiences and identify potential threats in social VR. Our results revealed a number of threats that teenagers may face in social VR. Some of the threats came from teenagers' experiences while others were observed by bystanders and parents. In this section, we reflect on our findings and further discuss how these findings shed light on nuanced forms of threats and social norms. Based on these findings, we also discuss the implications of designing safe and healthy social VR platforms as safe spaces.

5.1 Categorizing the Sources of Teenagers' Safety Threats

Our results suggested different types of safety threats that teenagers may face in social VR. Upon further examining these threats, we started to note the causes of these threats and grouped them into the following categories.

Discrepancies among the perceptions and experiences of teenagers, bystanders, and parents. Our data suggested teenagers, bystanders, and parents constantly held different opinions and/or experiences towards the same activities. Such mismatch may have led to some hidden threats which may not be obvious otherwise. For example, in the case of building connections with strangers in social VR, most of our teenager participants have normalized this action to be a fundamental aspect of social VR, yet parents and bystanders pointed out cases in which teenagers may face privacy and security risks due to the interaction with strangers. In the example of virtual grooming, our teenager participant built trust with the predator easily, yet bystanders who observed the situation tried to help the teenager but were refused, leading to greater risks of being harassed by the predator. In the example of physical aggression, our teenager participant who experienced physical aggression had disturbing feelings, yet their parent believed that it was an integral part of the game experience in social VR. When these discrepancies exist, teenagers would either not accept the help offered by others (since they believed that risks did not exist) or not ask for help when needed (since other stakeholders may not care about it). It became difficult to convince others to take proactive action and mitigate the potential risks.

Lack of social norms in social VR. Social norms, behaviors, and values in the physical world are shaped by socialization processes, cultural contexts, laws and policies, and broadly-acknowledged values. Similar to our physical world, social VR also represents a complicated social space that includes different types of users, events, and activities. Yet, the norms in our physical world may not necessarily translate to social VR environments. In fact, social VR did not seem to have established social norms that users follow to maintain a proper environment. For example, in the case of drinking alcohol in a bar, teenagers would not have access to an actual bar due to the age restriction. Yet, the lack of social norms in social VR made it possible for them to access the virtual bar and participate in activities, some of which might be inappropriate for teenagers (e.g., some teenagers were nudged to drink alcohol in real life). In the case of ERP, teenagers may also be exposed to sexual content (e.g., avatars with sexual symbols or signifies), which was against the established social norms in the physical world yet remained popular in social VR. We consider these types of threats as "hidden threats", which could be easily overlooked otherwise.

A challenge in identifying and defining social norms within social VR lies in its inherent anonymity. When users embody themselves through avatars, specific identity information (e.g., gender, age, and preferences) may be lost. However, the norms users are used to in the physical world are largely based on users' identity, and thus, are no longer effective in social VR. Future work should investigate human behaviors in social VR and help establish/identify appropriate social norms to ensure a healthy VR environment for teenagers.

Technological limitations and barriers. As social VR is evolving and not sufficiently mature, it also creates some technical limitations for users. For instance, moderators play an essential role in social VR, particularly in case users need help. However, moderators were not readily available in private rooms where safety threats were quite common. Instead, the responsibility of moderation is often left to the owner or creator of the private room who may not have the experience

to effectively manage the dynamic of the environment. As a result, these private rooms can inadvertently become safe havens for predators to engage in harmful activities, such as grooming, bullying, or exploitation of teenagers.

Additionally, VR devices also introduce limitations by providing an enclosed first-person experience only to the user. As such, our parent participants generally lacked participation in their children's VR activities. In fact, only a few parents in our study stated that they regularly played in VR with their children. Currently, social VR platforms do not support ad-hoc recording or checking history functions, making it difficult for teenagers to document their experiences and for parents to learn about these incidents. As such, this limitation further deepens the perception gap between teenagers and parents and may potentially cause more harm in the long run.

Finally, the immaturity of social VR ecosystems also contributes to teenagers' safety threats. For example, the process of avatar creation and customization also introduced further limitations. In our study, participants who wanted to customize their avatars needed to turn to third-party software or platforms (e.g., Unity, Blender). However, those platforms did not have proper guidelines or validation mechanisms to regulate the process. Social VR platforms also did not have power over these third-party platforms nor provided mechanisms to filter customized avatars other than some technical limitations (e.g., customized avatars cannot exceed certain sizes). As a result, users can freely create and utilize avatars to meet their individual needs which could potentially turn their avatars into vehicles of harassment (e.g., ERP).

5.2 Design Implications

Designing age-specific matching mechanisms for social VR. We propose the implementation of an age-matching system for social VR platforms, considering the significant usage of these platforms by teenagers and children. As highlighted in section 4.4, our participants expressed a strong preference for interacting with peers of a similar age. While some platforms offer junior accounts, the existing age verification system falls short of ensuring the accuracy of users' real age. We suggest that platforms consider implementing parental consent as a means of age verification. For instance, during the account creation process, the platform could send a link to the parents' phone that when clicked can allow them to sign a consent form. Moreover, while we acknowledge the possibility of users attempting to fake their age, additional ongoing monitoring measures can be put in place. These monitoring measures could involve the use of algorithms to detect suspicious behavior or inconsistencies in user profiles such as being reported multiple times or sending many unnecessary messages. By flagging potential discrepancies or anomalies in user activity, the platform can prompt further verification checks to ensure the accuracy of the user's age information.

Enable recording in social VR. We believe that it would be beneficial to incorporate a feature that allows users to share evidence with social VR moderators in case of unsafe and/or uncomfortable experiences. We propose the implementation of an "emergency button," similar to the screen recording functionality in the Zoom video conferencing software, to assist teenagers to request help when experiencing harassment, aggression, or other unsafe incidents in social VR. Activating this button would initiate the automatic audio and video recording of all activities within the user's vicinity, providing valuable evidence for future reference. It should be noted that to avoid abusing such a feature, the recorded media must be securely stored (ideally locally in VR headsets) by the social VR platforms and should be made exclusively accessible to the system moderators who can review them and take appropriate actions. Furthermore, the platforms should promptly notify moderators and parents when a user uses this feature to ensure their well-being.

Supporting non-tech-savvy parents and guiding children's social VR experiences. Concerns over the safety of teenagers in social VR have prompted many parents to seek further education in this field given their limited familiarity with the platform. To this end, one effective approach to address this need would involve updating existing safety education resources to include a dedicated VR component highlighting the potential risks and threats. Such materials could help raise awareness of harassment and sexual abuse issues in social VR among parents and their children and the same time equip children with the necessary knowledge to safely use social VR.

6 Conclusion

Social VR platforms have become increasingly popular in recent years among teenagers, yet safety issues such as harassment and sexual abuse continue to be significant concerns. This paper aims to investigate teenagers' experiences in social VR through three different perspectives, with a specific focus on harassment issues. Through an interview study with 8 teenagers, 9 bystanders, and 7 parents, we identified several threats for teenagers in social VR, including grooming and manipulation in private worlds. We also highlight new forms of harassment in social VR, such as Erotic Role-Playing and through phantom sense. Our findings provide a better understanding of the risks faced by teenagers in social VR and offer insights to design safer and more fulfilling experiences for them. We hope that our study contributes to the ongoing efforts to create safer social VR environments for teenagers.

7 Acknowledgment

We thank the anonymous reviewers and the shepherd for their invaluable feedback and participants for sharing their insights. This work is partially supported by a research gift from Meta.

References

[1] Samir Abou El-Seoud, Nadine Farag, and Gerard McKee. A Review on Non-Supervised Approaches for Cyberbullying Detection. *Int. J. Eng. Pedagog.*, 10(4), 2020.

[2] Mohammed Ali Al-Garadi, Mohammad Rashid Hussain, Nawsher Khan, Ghulam Murtaza, Henry Friday Nweke, Ihsan Ali, Ghulam Mujtaba, Haruna Chiroma, Hasan Ali Khattak, and Abdullah Gani. Predicting Cyberbullying on Social Media in the Big Data Era Using Machine Learning Algorithms: Review of Literature and Open Challenges. *IEEE Access*, 7, 2019.

[3] AltspaceVR. AltspaceVR to Sunset the Platform on March 10, 2023. https://altvr.com/sunset/, 2023.

[4] Chintan Amrit, Tim Paauw, Robin Aly, and Miha Lavric. Identifying child abuse through text mining and machine learning. *Expert Systems with Applications*, 88, 2017.

[5] Craig A. Anderson and Brad J. Bushman. Effects of Violent Video Games on Aggressive Behavior, Aggressive Cognition, Aggressive Affect, Physiological Arousal, and Prosocial Behavior: A Meta-Analytic Review of the Scientific Literature. *Psychological Science*, 12, 2001.

[6] Apple. Expanded Protections for Children. https://www.apple.com/child-safety/, 2023.

[7] Frank Biocca and Ben Delaney. Immersive Virtual Reality Technology. *Communication in the Age of Virtual Reality*, 15(32), 1995.

[8] Lindsay Blackwell, Nicole Ellison, Natasha Elliott-Deflo, and Raz Schwartz. Harassment in Social VR: Implications for Design. In *Proc. IEEE VR*, 2019.

[9] Doug A Bowman and Ryan P McMahan. Virtual Reality: How Much Immersion Is Enough? *Computer*, 40(7), 2007.

[10] Angela Browne and David Finkelhor. Impact of Child Sexual Abuse: A Review of the Research. *Psychological Bulletin*, 99(1), 1986.

[11] Larissa S Christensen, Dominique Moritz, and Ashley Pearson. Psychological Perspectives of Virtual Child Sexual Abuse Material. *Sexuality & Culture*, 25(4), 2021.

[12] Fast Company. If the metaverse is the future of social media, teens aren't convinced. https://www.fastcompany.com/90740073/if-the-metaverse-is-the-future-of-social-media-teens-arent-convinced, 2022.

[13] Lorrie Faith Cranor, Adam L Durity, Abigail Marsh, and Blase Ur. Parents' and Teens' Perspectives on Privacy In a Technology-Filled World. In *Proc. SOUPS*, 2014.

[14] Elmira Deldari, Diana Freed, and Yaxing Yao. Supporting a safe and healthy immersive environment for teenagers. *UMBC Student Collection*, 2022.

[15] Michele Di Capua, Emanuel Di Nardo, and Alfredo Petrosino. Unsupervised Cyber Bullying Detection in Social Networks. In *Proc. ICPR*, 2016.

[16] Dominic DiFranzo, Samuel Hardman Taylor, Francesca Kazerooni, Olivia D Wherry, and Natalya N Bazarova. Upstanding by Design: Bystander Intervention in Cyberbullying. In *Proc. CHI*, 2018.

[17] Matthew S Eastin and Robert P Griffiths. Beyond the Shooter Game: Examining Presence and Hostile Outcomes Among Male Game Players. *Communication Research*, 33(6), 2006.

[18] Pardis Emami-Naeini, Henry Dixon, Yuvraj Agarwal, and Lorrie Faith Cranor. Exploring How Privacy and Security Factor into IoT Device Purchase Behavior. In *Proc. CHI*, 2019.

[19] Marie Eneman, Alisdair A Gillespie, and C Stahl Bernd. Technology and Sexual Abuse: A Critical Review of an Internet Grooming Case. In *Proc. ICIS*, 2010.

[20] Samuel Farley, Iain Coyne, and Premilla D'Cruz. Cyberbullying at work: Understanding the influence of technology. *Concepts, Approaches and Methods*, 2021.

[21] Guo Freeman and Dane Acena. Hugging from A Distance: Building Interpersonal Relationships in Social Virtual Reality. In *Proc. ACM IMX*, 2021.

[22] Guo Freeman, Dane Acena, Nathan J McNeese, and Kelsea Schulenberg. Working Together Apart through Embodiment: Engaging in Everyday Collaborative Activities in Social Virtual Reality. In *Proc. GROUP*, 2022.

[23] Guo Freeman, Samaneh Zamanifard, Divine Maloney, and Dane Acena. Disturbing the Peace: Experiencing and Mitigating Emerging Harassment in Social Virtual Reality. In *Proc. CSCW1*, 2022.

[24] Guo Freeman, Samaneh Zamanifard, Divine Maloney, and Alexandra Adkins. My Body, My Avatar: How People Perceive Their Avatars in Social Virtual Reality. In *Extended Abstracts of CHI*, 2020.

[25] Philip Gillingham. Predictive Risk Modelling to Prevent Child Maltreatment and Other Adverse Outcomes for Service Users: Inside the 'Black Box' of Machine Learning. *The British Journal of Social Work*, 46(4), 2016.

[26] Google. Fighting child sexual abuse online. https://protectingchildren.google/, 2022.

[27] William R Graham Jr. Uncovering and Eliminating Child Pornography Rings on the Internet: Issues regarding and Avenues Facilitating Law Enforcement's Access to Wonderland. *L. Rev. MSU-DCL*, 2000.

[28] Hana Habib, Sarah Pearman, Jiamin Wang, Yixin Zou, Alessandro Acquisti, Lorrie Faith Cranor, Norman Sadeh, and Florian Schaub. "It's a scavenger hunt": Usability of Websites' Opt-Out and Data Deletion Choices. In *Proc. CHI*, 2020.

[29] Catherine Hamilton-Giachritsis, Elly Hanson, Helen Whittle, Filipa Alves-Costa, and Anthony Beech. Technology assisted child sexual abuse in the UK: Young people's views on the impact of online sexual abuse. *Children and Youth Services Review*, 119, 2020.

[30] Catherine Hamilton-Giachritsis, Elly Hanson, Helen Whittle, Filipa Alves-Costa, Andrea Pintos, Theo Metcalf, and Anthony Beech. Technology assisted child sexual abuse: Professionals' perceptions of risk and impact on children and young people. *Child abuse & neglect*, 119, 2021.

[31] Heidi Hartikainen, Afsaneh Razi, and Pamela Wisniewski. 'If You Care About Me, You'll Send Me a Pic'-Examining the Role of Peer Pressure in Adolescent Sexting. In *Companion Publication of CSCW*, 2021.

[32] Juan M Machimbarrena, Esther Calvete, Liria Fernández-González, Aitor Álvarez-Bardón, Lourdes Álvarez-Fernández, and Joaquín González-Cabrera. Internet Risks: An Overview of Victimization in Cyberbullying, Cyber Dating Abuse, Sexting, Online Grooming and Problematic Internet Use. *International Journal of Environmental Research and Public Health*, 15(11), 2018.

[33] Divine Maloney and Guo Freeman. Falling Asleep Together: What Makes Activities in Social Virtual Reality Meaningful to Users. In *Proc. CHI PLAY*, 2020.

[34] Divine Maloney, Guo Freeman, and Andrew Robb. A Virtual Space for All: Exploring Children's Experience in Social Virtual Reality. In *Proc. CHI PLAY*, 2020.

[35] Divine Maloney, Guo Freeman, and Andrew Robb. It Is Complicated: Interacting with Children in Social Virtual Reality. In *Proc. IEEE VRW*, 2020.

[36] Divine Maloney, Guo Freeman, and Andrew Robb. Stay Connected in An Immersive World: Why Teenagers Engage in Social Virtual Reality. In *Interaction Design and Children*, 2021.

[37] Divine Maloney, Guo Freeman, and Donghee Yvette Wohn. "Talking without A Voice": Understanding Non-Verbal Communication in Social Virtual Reality. In *Proc. CSCW2*, 2020.

[38] Divine Maloney, Samaneh Zamanifard, and Guo Freeman. Anonymity vs. Familiarity: Self-Disclosure and Privacy in Social Virtual Reality. In *Proc. VRST*, 2020.

[39] KF McCartan and Ruth McAlister. Mobile phone technology and sexual abuse. *Information & Communications Technology Law*, 21(3), 2012.

[40] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. Reliability and Inter-rater Reliability in Qualitative Research: Norms and Guidelines for CSCW and HCI Practice. In *Proc. CSCW*, 2019.

[41] Joshua McVeigh-Schultz, Anya Kolesnichenko, and Katherine Isbister. Shaping Pro-Social Interaction in VR: An Emerging Design Framework. In *Proc. CHI*, 2019.

[42] Joshua McVeigh-Schultz, Elena Márquez Segura, Nick Merrill, and Katherine Isbister. What's It Mean to "Be Social" in VR?: Mapping the Social VR Design Ecology. In *Proc. DIS*, 2018.

[43] Meta. Parent education hub. https://www.meta.com/quest/safety-center/parental-supervision/, 2023.

[44] Microsoft. PhotoDNA. https://www.microsoft.com/en-us/photodna, 2022.

[45] Kimberly J Mitchell, David Finkelhor, Lisa M Jones, and Janis Wolak. Use of Social Networking Sites in Online Sex Crimes Against Minors: An Examination of National Incidence and Means of Utilization. *Journal of Adolescent Health*, 47(2), 2010.

[46] Amgad Muneer and Suliman Mohamed Fati. A Comparative Analysis of Machine Learning Techniques for Cyberbullying Detection on Twitter. *Future Internet*, 12(11), 2020.

[47] Manja Nikolovska. The Internet as a Creator of a Criminal Mind and Child Vulnerabilities in the Cyber Grooming of Children. *JYU dissertations*, 2020.

[48] Zheng Qingxiao, Guo Freeman, and Andrew Robb. Understanding Safety Risks and Safety Design in Social VR Environments. In *Proc. CSCW1*, 2023.

[49] Ethel Quayle. Researching online child sexual exploitation and abuse: Are there links between online and offline vulnerabilities? *LSE Research Online Documents on Economics*, 2016.

[50] Ethel Quayle and Max Taylor. Social networking as a nexus for engagement and exploitation of young people. *Information Security Technical Report*, 16(2), 2011.

[51] Erin Romanchych. *Violent Video Gaming, Parent and Child Risk Factors, and Aggression in School-Age Children*. PhD thesis, University of Windsor (Canada), 2018.

[52] Rec Room. A Parent's Guide to Rec Room. https://recroom.com/parents-guide, 2023.

[53] Semiu Salawu, Yulan He, and Joanna Lumsden. Approaches to Automated Detection of Cyberbullying: A Survey. *IEEE Transactions on Affective Computing*, 11(1), 2017.

[54] Gökçe Nur Say, Zehra Babadağı, Koray Karabekiroğlu, Murat Yüce, and Seher Akbaş. Abuse Characteristics and Psychiatric Consequences Associated with Online Sexual Abuse. *Cyberpsychology, Behavior, and Social Networking*, 18(6), 2015.

[55] Ketaki Shriram and Raz Schwartz. All Are Welcome: Using VR Ethnography to Explore Harassment Behavior in Immersive Social Virtual Reality. In *Proc. IEEE VR*, 2017.

[56] Del Siegle. Cyberbullying and Sexting: Technology Abuses of the 21st Century. *Gifted Child Today*, 33(2), 2010.

[57] Mel Slater and Maria V Sanchez-Vives. Enhancing Our Lives with Immersive Virtual Reality. *Frontiers in Robotics and AI*, 3, 2016.

[58] Coen Teunissen and Sarah Napier. Child sexual abuse material and end-to-end encryption on social media platforms: An overview. *Trends and Issues in Crime and Criminal Justice*, 2022.

[59] VRChat. Community Guidelines. https://hello.vrchat.com/community-guidelines, 2019.

[60] Janis Wolak, Kimberly J Mitchell, and David Finkelhor. Does Online Harassment Constitute Bullying? An Exploration of Online Harassment by Known Peers and Online-Only Contacts. *Journal of Adolescent Health*, 41(6), 2007.

[61] Kexin Zhang, Elmira Deldari, Zhicong Lu, Yaxing Yao, and Yuhang Zhao. "It's Just Part of Me:" Understanding Avatar Diversity and Self-presentation of People with Disabilities in Social Virtual Reality. In *Proc. ASSETS*, 2022.

[62] Shikun Zhang, Yuanyuan Feng, Yaxing Yao, Lorrie Faith Cranor, and Norman Sadeh. How usable are ios app privacy labels? *UMBC Faculty Collection*, 2022.

[63] Xiaolu Zhang. Charging children with child pornography–Using the legal system to handle the problem of "sexting". *Computer Law & Security Review*, 26(3), 2010.

8 Appendix

8.1 Interview Protocol (Parent and Bystander)

8.1.1 Demographics

1. What gender do you identify yourself as?
2. How old are you?
3. What do you do for a living?
4. How many children do you have?
5. How old are they?
6. What types of devices do they have?
7. Which child(ren) uses VR?

8.1.2 Background

8. When did you buy your VR headset? Why?
   (a) What are the things you consider when buying a headset?
   (b) Have you ever tried it yourself?
   (c) Can you describe your experience?
9. Do your kids use VR? What do you think your child generally uses VR for?
   (a) How often do they use VR?
   (b) When was the last time your child used VR?
   (c) Do you know what they did?
10. In general, what do you think of VR?
    (a) Do you see any benefits of VR?
    (b) Do you have concerns about VR?
    (c) Have you ever heard of or experienced anything in VR that makes you frustrated?

8.1.3 Behaviors in VR

11. Have you ever heard of/used any social VR applications?
    (a) Can you provide some examples?
    (b) When was the last time you used *** (social VR apps)?
    (c) Can you walk us through what you did?
    (d) (Specifically, we want to follow up to see if they have ever interacted with anyone, like chat, talk, or other types of interaction) Did you interact with anyone?
    (e) If so, how? Did you approach them or the other way around?
    (f) What did you do? Why?
    (g) In this case, do you think the person you talked to is someone that can be trusted? Why?
12. (For parents) Do you know whether your child uses social VR or not?
13. What do you think about the idea of having your child (or teenagers) interact with other people in a virtual space? Would you support that?
14. From your perspective, what would be the reason why your child (teenagers) would like to interact with others in social VR?
15. In general, do you feel social VR is a safe place for your child? Why or why not?

8.1.4 Risks and Harms

16. (For parents) Do you know whether your child has any friends in social VR?
    (a) How did that start?
    (b) Are you supportive of these?
    (c) In fact, related to the last question, have you ever talked to your child regarding how to decide whether to interact with someone in social VR or not?
    (d) Do you have any rules or guidelines you follow?
17. (For parents) Has your child ever encountered any risks or harms when they use social VR?
    (a) How did you find out about it?
    (b) (If yes) Can you tell us a little bit about what happened? What did you do?
    (c) (If no) Have you ever seen any negative experiences happen to other people, like other kids or from other parents, or from the news?
    (d) Are there any signals you are looking for?
18. Have you ever encountered any negative experiences yourself or have you ever seen anything when you use it?

8.1.5 Safety by Design

19. Are you aware of any features or functions in social VR apps that can help ensure your safety when you are playing?
20. From your perspective, is there anything to be done to ensure the safety of the social VR space?
21. Now, imagine that you have a superpower that can be used to do anything. What changes would you make to the social VR apps you have used? (prompt: think from policy, technology, feature, design, etc.)

8.1.6 Wrap Up

22. Is there anything else you'd like to share?

8.2 Interview Protocol (Teenager)

8.2.1 Demographics

1. What gender do you identify yourself as?
2. How old are you?
3. What grade are you? Out of school? Working/college? (Depending on the age of the participant)

8.2.2 Background

4. Do you own any VR headsets?
5. What do you generally use VR for?
6. How often do you use VR?
    (a) When was the last time you used VR?
    (b) Can you tell us what you do?
7. In general, what do you think of VR?
    (a) Are there any cool factors?
    (b) Do you have any concerns?
8. Have you ever gone through anything in VR that makes you frustrated?

8.2.3 Behaviors in VR

9. Have you ever used any social VR applications?
   (a) Can you provide some examples?
   (b) When was the last time you used *** (social VR apps)?
   (c) Can you walk us through what you do?
   (d) (Specifically, we want to follow up to see if they have ever interacted with anyone, like chat, talk, or other types of interaction) Did you interact with anyone?
   (e) If so, how? Did you approach them or the other way around?
   (f) What did you do?
10. In general, why do you talk to other people in Social VR?
    (a) (If they have done that before in the prior case) So you mentioned that last time you talked to someone, is that for the same reason?
11. How did you decide who you can talk to and who you don't want to talk to?
    (a) (If they have done that before in the prior case) In that case, do you think the person you talked to is someone that can be trusted? Why?
12. In general, do you feel safe in social VR? Why or why not?

8.2.4 Risks and Harms

13. Have you ever encountered any negative experiences when you use social VR?
    (a) (If yes) Can you tell us a little bit about your experience, if you are comfortable? Please be assured that no one beyond our research team can hear what you said.
    (b) What did you do?
    (c) (If no) Have you ever seen any negative experiences happen to other people, like your friend or someone else in the social VR apps?
    (d) (If yes) Can you tell us a little bit about what happened?
14. Have you ever been approached by some other people in social VR apps, especially those you don't know?
    (a) (If yes) What did you approach you for? Can you talk a little bit about the scenario?
    (b) What did you do? Why did you do that?
    (c) How do you decide whether to respond to this person or not?
15. In fact, related to the last question, how do you decide whether to interact with someone in social VR or not?
    (a) Do you have any rules or guidelines you follow?
    (b) Are there any signals you are looking for?

8.2.5 Safety by Design

16. Are you aware of any features or functions in social VR apps that can help ensure your safety when you are playing?
17. From your perspective, is there anything to be done to ensure the safety of the social VR space?
18. Now, imagine that you have a superpower that can be used to do anything. What changes would you make to the social VR apps you have used? (Think from policy, technology, feature, design, etc.)

8.2.6 Wrap Up

19. Is there anything else you'd like to share with us?

Fight Fire with Fire: Hacktivists' Take on Social Media Misinformation

Filipo Sharevski
DePaul University

Benjamin Kessell
DePaul University

Abstract

In this study, we interviewed 22 prominent hacktivists to learn their take on the increased proliferation of misinformation on social media. We found that none of them welcomes the nefarious appropriation of trolling and memes for the purpose of political (counter)argumentation and dissemination of propaganda. True to the original *hacker* ethos, misinformation is seen as a threat to the democratic vision of the Internet, and as such, it must be confronted head on with tried hacktivism methods: deplatforming the "misinformers" and doxing their funding and recruitment. The majority of the hacktivists we interviewed recommended interventions for promoting misinformation literacy in addition to targeted hacking campaigns. We discuss the implications of these findings relative to the emergent recasting of hacktivism as a defense of a constructive and factual social media discourse.

1 Introduction

Steven Levy's portrayal of the hacker culture in his 1984 book *Hackers* largely remains the most influential reference to the public's general view of hackers [45, 67]. Recasting them as Robin Hood-style activists committed to a democratic vision of the Internet [101], Levy asserts that the hacker ethos embodies several sacrosanct postulates to the public good, notably that: (i) *all information should be free*, and (ii) *authority should be mistrusted and decentralization promoted* [67].

Later-day Internet hackers shifted towards an ideology oriented around autonomy in cyberspace. In this view, the Internet is seen as a politicized, public, information sharing space and as a valuable weapon against the neoliberal elites, who they see as responsible for economic and social disarray [40]. In other words, hacktivists took a front against the "*neoliberalism*" or the sociopolitical right-of-center positioning of individualized, market-based competition as the preferred governing principle for shaping human action in all areas of life – including the Internet – both at the individual and collective, societal levels [122]. Turning Internet activism into a form of socio-political resistance online [60] enabled a functional selection of issues that no longer necessitated lengthy preparations [77]. This, in turn, resulted in almost instant convergence and coordination of activities in response to the issues of interest. These campaigns in turn generated significant public visibility via coverage by mass media (e.g. television, newspapers, magazines, and radio) [49].

The Internet activism bifurcated to online campaigns concerned with the protection of the Internet as a relatively unregulated and unowned space (e.g. Anonymous, WikiLeaks, Snowden [23, 118, 120]) and online campaigns concerned with the protection of human rights and the environment (e.g. the Occupy movement, Arab Spring, Pirate Party [61, 84]). The former activism – or *hacktivism* – is often anonymous, performed in secret, and operates with a kind of impunity thus far afforded by networking technologies [121]. The latter activism – or *hashtag activism* – is usually public, openly leverages the Internet for political mobilization, operates primarily on the streets, and is subject to the dangers of crowd violence, harassment, and arbitrary arrest [104].

Hashtag activism historically utilized various technologies like petition websites (e.g. MoveOn.org for organizing political protests) or e-mail communication (e.g. Tea Party's campaign to reduce government spending and taxation) [18], but the advent of social media sites like Twitter, Facebook, and YouTube dramatically accelerated the self-organization and participation in the sociopolitical struggle (e.g. the #BlackLivesMatter and #SchoolStrike4Climate movements [37]). For hashtag activism there is a historical and ongoing essential dependence on social media [58]. The relationship between hacktivism and social media, however, is more complicated.

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.
USENIX Symposium on Usable Privacy and Security (SOUPS) 2023.
August 6–8, 2023, Anaheim, CA, USA

Hacktivists, in contrast, have hacked various technologies to defacing websites [102], broken into systems to "leak" private documents and "dox" individuals [118, 123], and have overwhelmed systems with traffic to cause a Denial-of-Service (DOS) [85]. Hacktivists' foray in social media mirrors these actions as campaigns were undertaken for hijacking/defacement of social media accounts (e.g., Anonymous's #OpKKK campaign [134]), doxing individuals on Twitter (e.g. the students of Covington High School [72]), and DoS Twitter topics (e.g. #IranTalks campaign [90]). But hacktivists also hacked the social media affordances for content amplification (e.g. StayWokeBot [39, 106]), early instances of trolling (e.g. Rickrolls [105]), and sharing memes (e.g. Lol Cats on 4chan [23]).

Despite the intuitive versatility of social media for such subversive operations, hacktivism became largely inactive on the mainstream platforms following some high profile run-ins of leading hacktivists with the legal authorities [55, 130]. The apparent absence of hacktivism created a vacuum where no one actively challenged the elites, defended freedom of expression, and appended the vision of democratic social media participation. It took little time, unfortunately, for this vacuum to be appropriated by state-sponsored actors hijacking the hacking playbook for actions aimed not just against the neoliberal elites but the entire social order [35]. Bot-enabled amplification aided political trolling and sharing of memes during the Brexit campaign in the UK [26] and the 2016 elections in the US [11]. The crucial difference in these instances was that the amplified memes and trolling were not pranks but damaging fake news, emotionally-charged memes, and conspiracy theories that instead of unifying the social media crowds for a cause, divided them in opposition camps that were pitted against each other [115].

In response to such a large-scale disruption on the social media landscape, one would have plausibly expected for hacktivists to retaliate, confront, expose, or counter-hack the state-sponsored "trolls" [141]. Misinformation, back to the Levy's depiction of hacker's ethics [67], runs counter the first postulate (i) *all information should be free* because it undermines the basic utility of information as a public good (i.e. truth and facts do not dwindle in supply as more people "consume" them and truth and facts are available to all people in a society) [34]. Misinformation also runs counter the second postulate (ii) *authority should be mistrusted and decentralization promoted* because it is promulgated by a state-sponsored "shadow authority," as evidence confirms in the aftermath of the Brexit and the 2016 US elections [50, 75, 140]. Surprisingly, the hacktivists never struck back [12], though they clearly possessed the capabilities to do so, as evidenced in the Anonymous's #OpISIS campaign, for instance, where the collective flagged about 101,000 Twitter accounts attributed to the Islamic-State [51].

The absence of response to misinformation on social media by the hacktivist community seemed quite perplexing and, in our opinion, worthy of in-depth inquiry with active "hackers" that still operate in the spirit of the Levy's code of ethics [67]. Through personal connections and snowball sampling, we identified 22 prominent hacktivists and conducted hour-long interviews with each of them to learn their take on the misinformation ecosystem, on responses to falsehoods on social media, and on the way misinformation impacts and shapes the hacktivists' agenda in the future. We found a consensus among the hacktivists against the present forms of misinformation as an ammunition for political counter(argumentation) and external propaganda. They recommended actions to deplatform, dox, and expose every "misinformer" that is believed to pollute the social media discourse, and suggested ways to improve the general misinformation literacy among users in addition to these targeted operations.

To situate our study in the intersection between the hacktivist counter-culture and the rise of misinformation on platforms, we review the interplay between Internet activism, social media, and false information in Section 2. We look in the broader context of misinformation in Section 3 to highlight the pressing need of (hack)tivism action to reclaim the social media space true to Levy's vision of Internet as an information exchange to the public good. In Section 4 we outline our research design and methodology. Sections 5, 6, and 7 expand on our findings and we discuss the implications of the hackers' disposition to social media misinformation in Section 8. Finally, Section 9 concludes the paper.

2 Internet Activism and Social Media

2.1 Hashtag Activism

Online social media activism – or *slacktivism*, *clicktivism* – emerged on popular platforms as a repertoire of low-risk, low-cost expressive activities for advocacy groups' agenda setting and political participation [103]. Social media users participated in petitions, changed personal avatars, added picture filters in support of a cause, and simply "liked" posts as an act of participation [43]. Slacktivists quickly realized they could use virality as a distinctive social media affordance to their advantage and move to use hashtags as the main drivers of mobilization, raising awareness, and demanding sociopolitical change. The practice of *hashtag activism* was instrumental for the success of social movements like #metoo, #takeaknee, and #BlackLivesMatter, allowing for visibility, expression of solidarity, and statement of victimhood [119]. This success, in turn, inspired a plethora of other movements advocating for health, human rights, social justice, and environmental issues across all social media platforms as a trend that remains active and prominent across online public discourse [54].

The advent of the hashtag activism, however noble, had to deal with the obvious threat of *hashtag hijacking*, or the appropriation of viral hashtags as a vehicle to inject contrary perspectives into the discourse [132]. This "hack" against

Internet activism is not just adding noise or attempting to result in a DoS, but is also used to disseminate hateful narratives and dilute the campaign itself (e.g. the hijacking of the #metoo hashtag [71]). Another similar threat is *hashtag co-opting*, or the contentious co-opting of the rhetoric of popular social movements (e.g. #HeterosexualPrideDay campaign co-opting the language of the mainstream LGBT movement [8]). Equally threatening is *counter hashtagging*, which concocts similar hashtags to garner opposition to well-established movements (e.g. #BlueLivesMatter countermovement to police reform in reaction to #BlackLivesMatter [63]). These antagonistic appropriations of social media virality enable political extremism to creep in the public discourse and embroil users in an emotionally-charged participation [99].

In an age of emerging social media polarization, it was a matter of time before fake news, offensive memes, and conspiracy theories would be weaponized against hashtag activism (e.g. the proliferation of fake news in the #Gunreformnow vs #NRA Twitter battle [20]). What was initially expected to remain on the fringes of the mainstream hashtag activism [36], quickly turned into information disorder on a mass scale. Today hashtag hijacking and co-opting develops *in parallel* with activism campaigns, feeding from and perpetuating an ecosystem of false and unverified information. This emotionally-charged participation has manifested within a global health panic (e.g. #FlattenTheCurve hashtag hijacking for COVID-19 misinformation [29]) and moral panic (e.g. the QAnon's co-opting of #SaveTheChildren hashtag [87]) in addition to the already growing political panic [89].

2.2 Hacktivism

Hacktivism was a term that "Omega," a member of the Texas-based computer-hacking group *Cult of the Dead Cow* (cDc) coined in 1996 in an email to the cDc listserv [78]. Characterized with the increasingly political ethos of hacking-for-cause, hacktivists primarily leveraged technology to advance human rights and protect the free flow of information in campaigns against the UK, US, and Chinese governments, as well as the UN [92]. In as much as hackers individually roamed the Internet, socialization was increasingly desired as many of them needed to establish a strong hacktivist network. Hacktivists' penchant for humorous memes (LOLCats) and gag hyperlinks (Rickrolls) [95] attracted an army of hackers to Christopher Poole's `4chan.org` social media website, setting the stage for the notorious hacktivist collective Anonymous [78].

While these hacktivists never displayed a predictable trajectory in their cyberoperations and political program [23], they narrowly utilized social media for self-promotion – announcing operations with an #Op prefixed hashtags [12] – and furthering relationships with other Internet activists. Anonymous cried foul on Twitter when WikiLeaks puts millions of its documents behind a pay wall [42], but also launched operation #Ferguson which doxed the St. Louis County police chief daughter's information in response to the shooting of the black teenager Michael Brown [10]. Hacktivists, in solidarity to the Arab spring uprisings, sent a care package composed of security tools and tactical advice though downplayed the touted "Twitter Revolution" [23].

True to their credo for utilizing Internet technologies against oppression, including social media, hacktivists launched the #OpKKK in support of #BlackLivesMatter protesters in Ferguson, Missouri to "unhood approximately 1000 Ku Klux Klan members" by gaining unauthorized access to a KKK Twitter account [134]. After a several years hiatus, perhaps due to arrests of some of the leading Anonymous hacktivists, the group resurfaced during the 2020 #BlackLivesMatter protests in response to the killing of George Floyd [56]. This time, in addition to leaking a 269 gigabyte trove of confidential police data (dubbed *BlueLeaks* [66]), the hacktivists launched social bot operations to amplify the online support for #BLM and criticize police actions.

Anonymous-affiliated hacktivists also utilized Internet technologies in the context of cyberwarfare. For example, the #OpIsis operation, which collated and published lists of tens of thousands of Twitter accounts that purportedly belonged to members of ISIS or its sympathizers, was launched in response to the terrorist attacks in France in 2015 [80]. Here, in addition to the identification efforts, hacktivists also waged a meme war and called for a "Troll ISIS Day" to provoke and disrupt ISIS-supported social media [79]. In early 2022 the Anonymous group took to Twitter to declare a "cyber war" to Russia in response to the Ukrainian invasion, launching DoS attacks against Russian's Federal Security Service's website and hacking Russian streaming services to broadcast war videos from Ukraine [108].

3 Internet Activism and Misinformation

3.1 Grassroots Misinformation Operations

Hacktivists, perhaps inadvertently, authored or gave popularity to the most utilized primitives for creating, propagating, amplifying, and disseminating misinformation - *trolling* and *memes*. This negative externality is unfortunate as trolling and memes were initially used by Anonymous against what they perceived a "misinformation campaign" by the Church of Scientology [78]. The "anon" members on `4chan.org` practically *hijacked* the term "troll" – initially meaning provoking others for mutual enjoyment – to abusing others for members' own enjoyment by posting upsetting or shocking content (usually on the /b/ channel of `4chan.org` [23]), harassing users (e.g. mocking funeral websites [14]), and spreading rumors [64]. What Anonymous did for the "lulz" (a brand of enjoyment etymologically derived from laughing-out-loud (lol)), nonetheless, showed the ease with which one could exploit the Internet technologies to be impolite, aggressive, disruptive, and manipulative to users' emotional states [23].

Trolling initially came in textual format as comments to posts, bulletin boards, and websites "deindividualized" people's lived experience for the "lulz" [14]. Gradually, hacktivists popularized a multimedia format of trolling or "memes," where textual commentary is superimposed over well-known imagery, typically representing different forms of power, such as political leaders, the police, and celebrities [79]. Memes, perhaps, were the actual rite of passage to true hacktivism – moving away from the early LOLCats – as they seek to deconstruct the power represented, contest censorship, and provide political commentary [91]. Memes as content were put to hacktivist use *en masse* in operations like "Troll ISIS day," where Anonymous proliferated memes with rubber-duck heads or rainbow stripes to ridicule ISIS propaganda imagery and disinformation narratives on Twitter [79]. Spread together with satirizing hashtags (e.g. #Daeshbags), the trolling memes achieved a cultural virality that brought hacktivism into the mainstream discourse online [96]. What the hacktivists did with the memes nonetheless, showed the ease with which anyone could disrupt, challenge, reimagine, and appropriate new political contexts by harnessing the virality and visibility of content spread on social media [88].

3.2 Mainstream Misinformation Operations

The hacktivists' playbook of trolling and meme dissent, though initially targeted *against* misinformation, was skillfully appropriated for the purposes of crafting and disseminating misinformation from 2014 onward, coinciding with the period of hacktivist inactivity [12]. This playbook alone was at first insufficient to achieve widespread political disruption, as it necessitated a support network of many accounts to gain traction. But the "appropriators" – privy to prior campaigns of disinformation and with the support of nation-state governments [117] – did not need to look further than the "sock puppet" accounts which were already utilized for spreading political falsehoods (e.g., Martha Coackey's "twitter bomb" disinformation campaign [89]). Having all the ingredients necessary to exploit the virality of social media and users' familiarity with emotionally-charged discourse, the "appropriators" established *troll farms* in the lead up to the UK's Brexit campaign and the 2016 US elections [75, 141].

The "army" behind the troll farms were particularly clever to integrate their social bots with "sock puppet" accounts that imitated ordinary users to systematically micro-target different audiences, foster antagonism, and undermine trust in information intermediaries [7]. Playing both sides in the emotionally-charged discourse already unfolding on social media, the troll farms posed as authentic, culturally competent personas (e.g. the so-called "Jenna Abrams" account [136]), and as vocal supporters of hashtag activism (counter) movements (e.g. BlackToLive in #BlackLivesMatter and SouthLoneStar in #BlueLivesMatter [124]). They also appropriated hashtag hijacking (e.g., #elections2016 and #ImVotingBecause tagging of quotes about Donald Trump and against Hillary Clinton [4]), hashtag co-opting (e.g. #BlackGunsMatter and #syrianlivesmatter [31]), and counter hashtagging (e.g. #NoDAPL against the Dakota Access Pipeline [47]). The troll farms even had the audacity to impersonate Anonymous themselves (e.g. the @_anonymous_news impersonation of the "Your Anonymous News" twitter account [22]).

The "meme game" of the troll farms was equally sophisticated and added to the initial success of their operations [86]. Trolls tested the waters around war-related memes regarding the opposition/support of the conflict in Syria [31], capitalized on both meme trolling and Internet activism to spread political memes through their fabricated Blacktivist social media accounts and co-opted Wikileaks in exploiting the leak of sensitive documents from the Democratic National Committee (DNC) [73]. Memes were also used to amplify conspiracies (e.g. QAnon, Pizzagate, and the murder of Seth Rich [138]), Texas secessionism (e.g. if Brexit why not #Texit [52]), and direct attacks (e.g. crooked Hillary [48]).

While the initial campaigns of the troll farms have been tracked, exposed, and brought into attention [31, 48], social media discourse has not recovered from this watershed period of meme and trolling appropriation for the purposes of conducting large-scale information operations [115]. Worse, the troll farm brand of political dissent was adopted by populist accounts keen on disseminating misinformation beyond just politics [53]. The trolling pandemonium spilled out of control with the COVID-19 pandemic as rumors, conspiracy theories, fake news, and out-of-context spins plagued the social media by hijacking the dominant hashtags like #COVID19, #coronavirus or #DoctorsSpeakUp [15], co-opting hashtags like #plandemic [62] and counter hash tagging with hashtags like #COVIDIOT [114]. Memes were distributed in conjunction with deepfake videos on platforms like YouTube [100] and TikTok [9] as well as blatant fake news on alt-platforms like Gab [21] to effectively reach a self-perpetuating bedlam of misinformation Internet counter-activism.

4 Hacktivism and Misinformation

In a radical state of ravaging misinformation campaigns on social media with no end in sight, one could wonder what the original activists on the Internet have to say in response. The unravelling of falsehoods is clearly a serious threat to the democratic vision of the Internet [101], as misinformation facilitated the rise of non-democratic communities contesting even factual knowledge and science (e.g. anti-vaxers, climate change deniers, etc. [133]). Hacktivists, as we have seen in Section 2, have fiercely opposed early misinformation campaigns in the past, but their means to do so were later "hijacked" for the purposes of mass misinformation production (i.e."disinformation" when spread with *intent* to deceive). One could attribute the paucity of hacktivists' involvement in the passing of the techno-liberal order of the Internet as the

rise of partisan-divided trust in facts and the politicization of science were already underway [38], but that alone is not a sufficient showstopper for action.

Regardless of any new Internet order, there is a reasonable expectation that one should still act upon the Levy's sacrosanct postulates [67], even if operating within an ecosystem polluted with misinformation. In addition to the public good arguments, misinformation is in conflict with the first (i) *all information should be free* postulate as it creates "information disorder" that, by the token of catalyzing polarization and emotionally-charged participation online, gives even more power to the neoliberal elites for perpetuating the economic and social (media) disarray [27]. Misinformation also conflicts with the second postulate (ii) *authority should be mistrusted, and decentralization promoted* as it stands in the way of independent truth discovery and dissemination online [69]. Should the new brand of reprehensible misinformation, therefore, be on the top of the hacktivists' agenda already?

4.1 Research Questions

To explore the gap in response to mass misinformation, we invited prominent hacktivists to address these questions:

- **RQ1:** How do contemporary hacktivists conceptualize the social media misinformation ecosystem?
- **RQ2:** What actions do hacktivists deem appropriate in response to misinformation on social media?
- **RQ3:** In what directions do the hacktivists see the misinformation ecosystem evolving toward in the future?

4.2 Sample

Our study was approved by the Institutional Review Board (IRB) of our institution before we invited, through personal contacts, and snowball sampling the hacktivists for a virtual interview session during 2022 and early 2023 with open-ended questions, listed in the Appendix. We sampled a population who were 18 years or older, from the United States, and that is an active contributor in the hacktivist community. As "active contributors" in the hacktivists space, our participants stated they are concerned with challenging online far-right extremism, help tracking criminals, and uncovering foreign countries' information operations. All of them were active in hacktivism prior to 2014 (and we conducted Open Source Intelligence (OSINT) investigations to verify that there is no evidence of them engaging in harmful or criminal activities in the past). The participants in our sample identified themselves using Levy's hacker ethos [67] and maintain presence on mainstream social media (Twitter, Facebook) and chat-based communities (Discord, Matrix). We used Zoom to conduct the interviews and allowed participants to choose whether to share a video feed or not. Every interview was recorded, stored in a secure server, and manually transcribed. We communicated with each interviewee to obtain final approval prior to starting the qualitative analysis.

Overall, our final sample contained 22 participants, all of which agreed to participate voluntarily. Gender demographics are given in Table 1. We made a deliberate attempt to produce a sample that is not a male-only or male-dominated, as previous studies indicate that the hacktivist community is imbalanced in regards gender [126]. Participants identities were not anonymous to us as researchers, but we deliberately suppress identifying statistics and potentially identifiable information in the reporting of our results to preserve participant anonymity to the general population, as a condition for their participation. In some cases, we used a direct censoring of names in citing participants' responses. We allowed the participants to skip any question they were uncomfortable answering. Each interview took around an hour to complete.

Table 1: Sample Demographic Distribution

Gender		
Female	**Male**	**Non-Binary**
8 (36.4%)	13 (59.1%)	1 (4.5%)

4.3 Methods and Instrumentation

To ensure validity to the task of conceptualizing misinformation, we introduced the participants to the generalized definition of social media misinformation from [135]. This also helped avoid confusion between past trolling and memes "for the lulz" and present information operations involving rumors, conspiracy theories, hoaxes, and clickbait. The hacktivists in our sample were invited to speak about their profiles, activity, and agendas online, before we asked their take on misinformation on social media. The qualitative responses were coded and categorized in respect to the following seven themes: a) antecedents to misinformation; b) mental models of misinformation; c) countering misinformation through leaking, doxing, and deplatforming; d) anti-misinformation operations (referred to as "ops") ; e) counter-misinformation tactics; f) misinformation literacy; and g) misinformation hacktivism.

Two independent researchers analyzed the interview transcriptions, achieving a strong level of inter-coder agreement (Cohen's $\kappa = .82$). We utilized a thematic analysis methodology to identify the aforementioned themes that naturally emerged from the responses in our sample. These themes were summarized to describe the conceptualization of, response to, and evolution of misinformation in the view of the contemporary hacktivists we sampled. In reporting the results, we prioritized verbatim quotation of participants' answers (emphasized and quoted in "*italics*") but omit reference to participant number in the sample to preserve their anonymity.

4.4 Hacktivists' Profiles

The hacktivists in our sample, true to the original ethos, represent the voice for advocacy and contemporary policy discussion. While they did not disclose their current operations, several of them hinted they are involved in tracking the rise of the far-right extremism, cybercriminals, as well as the information warfare part of the Ukraine invasion. A few of the hacktivists reported an agenda which comprised of leaking documents from companies and nation-state agencies as manifestation of their information freedom advocacy. Few of the hacktivists explicitly mentioned they still create and disseminate memes and participate in the "old school" trolling. And several of the hacktivists did actual *hacking* as in analyzing security problems (e.g. ransomware) and providing free tools for helping ordinary Internet users fend off related threats.

The majority of the hacktivists noted they have been active for a long time, being brought into the world of computers in childhood or early adolescence. Some of them resorted to hacktivism as a way to protect themselves against online bullies and some of them in response to state-sponsored offensive operations online, notably campaigns attributed to China and Russia. Several of them started with hacking operating systems to enable unrestricted access to games and/or bypass parental controls. While most of the participants in our sample cited curiosity as their driver to enter the "hacktivist conglomerate" and keep on hacking, there were many who voiced a strong support for cybersecurity education activism.

5 Misinformation Conceptualization

Evidence shows that social media users use multiple models to conceptualize misinformation – not just the traditional model that narrowly focuses on the fallacious nature of the information [113]. Beyond just fake news, misinformation is equally conceptualized as form of *political (counter)argumentation* where facts do selectively appear in alternative narratives relative to political and ideological contexts, often taken *out-of-context* with speculative intentions. Misinformation is also seen as *external propaganda* that includes *manufactured* facts and factoids disseminated and amplified online with the intention to create division. Given the radical transformation of the trolling and memes over time, our first research question aimed to learn the hacktivists' take on these competing conceptualizations amongst ordinary social media users.

5.1 Antecedents to Misinformation

The participants in our sample agree that trolling and meme dissemination has been hijacked for nefarious purposes, pointing out that they are not surprised about the current misinformation proliferation on the Internet. One participant summarized this evolution through first account experience:

> *I remember using sock puppet accounts way back in the early 2000s running forum raids as a* [redacted]*, specifically to run/post misinformation on other forums online. It was mostly for laughs, but we were massive monsters in those days. The only real major difference is these days is that the sock puppets are automated and put in action for keeping people tribalistic and resistant to opposing views.*

The use of "*sock puppets for running forum raids in the old days of hacktivism*," unfortunately, was not a serious enough threat for social media companies to implement "*strict policies of who and how can participate in the public discourses early on*" and counter to their business model of "*monetizing every possible engagement on their platforms,*" in the view of our participants and true to their innate resistance against the neoliberal appropriation of Internet freedoms.

Mainstream social media companies were accused of being the direct enablers of the "information disorder" as their models of engagement pushed "*less educational content the more an issue was important and demanded action*." This disorder played in the hands of the neoliberal elites and media outlets run by "*billionaires detached from reality to gain further control over public spaces*" as one of the participants put it. In the view of our participants, misinformation "*has always been there*" and pointed to the combination of "*self-proclamation of expertise online, cultivating followers, and playing on confirmation bias*" as the recipe the very hacktivists showed it works well in seeding misinformation:

> "*For example, look at the* [redacted]*. This person said they were a founding member of Anonymous and lots of people believed them. The person has spoken at conferences about it and even got jobs because of it. Literally dig slightly into that and it's clear that no one in the Anonymous community can vouch for the person and there's no evidence of them being linked. So, people are just too lazy to check stuff out because this person is kinda selling a story that fits with what they think so it must be true*."

5.2 Mental Models of Misinformation

The predominant mental model of misinformation amongst the hacktivists in our sample was *political (counter)argumentation*, where misinformation is disseminated for the sake of furthering a political argument or agenda [113]. In the original version of trolling and meme sharing the misinformation was seen as an alternative expression of disagreement, revolt, or ridicule without any context, but contemporary trolling and memes enter into the political context as content ready-made for the expression of political attitudes [94]. Despite fact checking being widely

available (and even suggested to users when content is moderated on social media [112]), the political appropriation of misinformation thrives because "*people won't fact check things and perpetuate them as long as these things align with their political ideology*." The reason why most social media users "*fall for misinformation*," in the view of our participants, is "*plain ignorance and stubbornness to hear anything contrary to their own political opinions*." One participant offered the following genesis of the misinformation problem:

> "*I think that people have learned that spreading disinformation through social media, Twitter for example, it's one of the best ways to get a word out. Twitter readers won't fact check things, especially if it aligns along with a political ideology people are passionate about so this word gets effectively to them. They'll believe whatever you tell them, and I think this is because there's a serious lack of, at least in the US, critical thinking education in schools.*"

In the view of the majority participants,"*both sides of the political spectrum spread misinformation and it further enables polarization*." While they acknowledge that "*the misinformation on social media is often identified with right-wing opinions*," participants recognize that "*we overuse the terms misinformation and disinformation to describe anything that is not a leftist opinion or fact.*" They point to the misinformation "stickiness" where the repeated exposure to speculative and false statements make them appear truthful [68], becoming the main theme of social media discourse. For example, one participant pointed out the Hunter Biden laptop saga [46]:

> "*It's usually the outrageous political claims that attract a lot of attention and people want a proof of concept, right? For example, take the bold claims aligned with the political message behind the Biden's laptop. Maybe there was a laptop but it's been politically disinfoed [sic] to death, to the point that the laptop leaks are irrelevant and can't be trusted as an evidence. These politicized things require a deeper dive into the actual truth as bold claims require bold evidence, but that's often missing so disinformation naturally creeps in.*"

Misinformation as political counter(argumentation) conflicts with the *all information should be free* postulate, which in turn forces mainstream social media platforms to "*restrict the flow of information.*" Misinformation, in the view of one of the participants, should not be restricted because "*people are entitled to see both sides of a proverbial political coin so the platforms must allow them to do so, otherwise by only showing heads or tails people will speculate about what's on the other side and assume the worst.*" The restriction of information on platforms conflicts with the *mistrust of authority and promote decentralization* hacker postulate because it allows "*the elites to define what constitutes 'truth' alone*," according to one participant. It also forces "*people to become rather tribalistic and a priori suspicious of people with different views.*" The "political tribalism" on social media [3], in turn, makes it "*easier to demonize people with different opinions and political attitudes and avoid scrutinizing the like-minded ones,*" playing directly in the hands of the "misinformers."

As for the "misinformers", our participants identified the state-sponsored "appropriators" that hijacked the original hacktivist playbook to spread *external propaganda* on social media. That other countries promulgated disinformation was not a news to the hacktivists (e.g. "*Russia has always been really good at it*"), but instead what surprised them was the "*audacity and the sophistication*" in utilizing trolling and memes on such a massive scale [140]. Reflecting on this shift in online operations, one participant believes that "*disinfo operations and hacking our intellectual property is all these other countries are left with because they can't beat US militarily or economically*." Not necessarily neoliberal, but nonetheless authoritarian, the elites behind the external propaganda in equal degree conflict with the *mistrust of authority and promote decentralization* hacker postulate because they are behind a "*blatant effort to control the social media turf and the mass of population spending their time there*", per one of the participants. The external propaganda nature of disinformation also conflicts with the *all information should be free* hacker postulate in the view of the hackers in our sample because "*overshadows and complicated an access to other more factual or useful information*."

6 Active Countering of Misinformation

Literature on misinformation focuses on helping the social media users discern falsehoods with strategies for "prebunking" (i.e. forewarning and preemptive refutation of the falsehoods [70]) or "debunking" (i.e. providing users verifiable corrections of the falsehoods from credible sources to break the illusion of truth [33, 97]). Algorithmic moderation tools are also available to mainstream social media platforms (the alternative ones do not deem misinformation as a problem [111]). These tools leverage natural language processing, image analysis, or metadata to detect trolling and memes [52,53,128]. Platforms have the option for "soft" moderation (by either obscuring trolling and memes with warnings covers or attaching warning labels [112,131]) and "hard" moderation (removing or suspending misinformer accounts [65]). None of these solutions, however fends off troll farms and meme disseminators effectively. Our second research question, therefore, sought to query hacktivists for their thoughts on countering this development.

6.1 Leaking, Doxing, and Deplatforming

The suspension of user acounts by social media platforms for breach of their code of conduct is referred to as "deplatforming" [2]. In the context of hacktivism, it takes a broader meaning, as hacktivists do investigative work that entails leaking and doxing but also confrontation with the misinformers that, in their subjective view, contradict the vision of a democratic Internet. For example, hacktivists did a massive API scraping of the alt-platform Parler to leak data that tied users to the Capitol Riots and the QAnon conspiracy [98], which in turn resulted in a massive account deplatforming on Twitter [17]. These activities spurred operations to confront and expose the QAnon conspirators on social media (e.g. @QAnonAnonymous [24]), amongst which some of our hacktivists have a direct role in "*dismantling the Qanon infrastructure*."

This deplatforming targeted political misinformation campaigns where our hacktivists "*compiled and leaked dossiers on individuals spreading hateful propaganda and those who seek to sow the seeds of violence*" on social media. These operations were targeted both on "*individual spreaders, nation-states, even companies with murky records*." Several mentioned their direct operations for exposing disinformation relative to the "*Ukrainian conflict*," praising the work of the Ukrainian IT Army outfit for dispelling the myth that Ukraine is committing genocide against Russians in the Donbas region [25]. Hacktivists were dedicated to "*doxing companies and governmental agencies in response to the political meddling in the US internal affairs from places like Russia, Iran, and China*." Misinformation "*sanctioned by the governments*" was targeted by the hacktivists in attempts to deplatform prominent "*disinformation front agents on social media, like* [redacted]*, for example*."

Leaks and doxing were equally utilized for misinformation beyond political counter(argumentation) and external propaganda. One of the participants has dedicated considerable time on exposing cryptocurrency scammers on social media and elsewhere, deeming the feeling of it as "*better than sex*." Another pushed back against criminal misinformation by doxing "*bullies, liars, and fraudsters*" and one "*anti-cancel culture in case of minors*" hacktivist noted that they "*successfully deplatformed major participants in hate campaigns and stalking of minors*" on social media. Another focused on leaking personal details about predators on social media who spread misinformation to cover their sexual harassment and cyberstalking towards women:

> *I've called it sometimes when I notice it on Twitter or elsewhere. I've exposed threat actors after tracing their activity and positively identifying them. Unfortunately, this is somewhat of a Bushido violation amongst fellow hackers but I am not concerned with such things. Some of these clowns have it coming to them. There was one person who went by the handle* [redacted] *who had been sexually harassing women, cyberstalking them, creating several sock puppet accounts, and just generally being a real nuisance in the community. Well, I doxed that person's real name on Twitter but I didn't post their address. This person thought he had cleaned up his tracks online being an 'infosec' professional but he underestimated someone like myself with a technical OSINT background. I easily found their information in an old resume on the Way Back Time Machine internet archive and posted their name. Some fellow 'infosec' pros didn't appreciate that I did so but honestly, the person had it coming and I don't regret it. I was careful about what I shared so no physical harm came to the them.*

6.2 Anti-Misinformation Operations

The hacktivists in our sample engaged in misinformation saturation operations, true to the their commitment to fight misinformation with more information. One of the hacktivists stated that it is "*expected from the hacktivist community to combat misinformation in such a way*" and noted that "*it is the sole reason they maintain a Twitter account.*" Another one seconded this posture noting that "*it is frustrating to see misinformation from others and other creators but that is the main reason I continue to post on TikTok.*" In the words of one participant, "*there is more ideological aspect of it when I am fighting disinformation*," directly invoking the mission of the true hacktivists to become reflexively "*loud and determined*" to speak true information in response to the "*general assholery of misinformation on internet.*"

Partaking in operation #NAFO (North Atlantic Fellas Organization) dedicated to countering Russian propaganda and disinformation in Ukraine by weaponizing memes [107], our participants materialized a combination of saturation and doxing to "*curtail misinformers' ability to gain followers.*" They extended their work to counter "*extremists and fascists and their toxic conspiracy theories*" by disrupting their funding and deplatforming prominent followers, true to the spirit of the "Antifa" hacktivist counterculture [137]. In a similar vein, one of the hacktivists proclaimed that they "*greatly contributed in the #OpJane operation*." #OpJane is the latest operation launched by Anonymous against Texas for enacting the anti-abortion Bill 8 that allows "abortion bounty" for anyone who anyone who reports abortion in the state of Texas [41]. Interestingly, in the announcement of the operation, Anonymous calls for "fighting misinformation with enough plausible and difficult to disprove misinformation" to make any data these bounty hunters gather as useless [6].

7 Misinformation Evolution

As there is virtually no cost to disseminating misinformation [89], it is unlikely that the online discourse will rid itself

of the alternative narrative plague any time soon. Whether this gloomy prediction will eventually materialize [81], or whether new technologies will improve the public's ability to judge the quality and veracity of content [5], remains an open question. Because hacktivists are nonetheless stakeholders in resolving this issue, our third research question aimed to learn their thoughts about how online spaces will fare with trolling, memes, and falsehoods in the near future.

7.1 Counter-Misinformation Tactics

The hacktivists in our sample unanimously posit that "*it is hard for social media platforms to keep up with removing it, so people stepping in to help is going to be of critical importance*" for preserving a healthy discourse. The mobilization for "*justice and truth as a cause*" is important not just for curbing misinformation but also in "*reclaiming information back from the political hold.*" To help "*expose misinformation charlatans,*" hacktivists call for maintaining a code of conduct where "*no leak, doxing, or exposure action should cause anyone else harm (physical, reputation, mental).*" To begin with, one participants reckons we should do the following:

> "*If one is a disinformation actor and they're acting aggressively I feel like you have to respond in a similar measure, in this case. I identify what their weaknesses are, what is it that's going to trigger them? Trying to get their accounts to get shut down, trying to get them to react in a way that will expose them. That's something I think is fair, as long as you're doing it [via] legal means. Getting open-source information about the individuals, exposing them, I think that's totally fair. Disinformation actors always try to be anonymous, of course, but what is the intent of that? Being anonymous allows you to act with impunity to do these really nasty things? Whereas all of a sudden if the tables are reversed and a disinformation actor is exposed, now I feel like we're teaching them a lesson. I guess it is sort of vigilantism, but in certain cases it's warranted. And one thing we got to do is got stop treating disinformation as freedom of speech. It's one thing to think you can say what you want, but that shouldn't shelter you from the consequences.*"

As misinformers usually use the anonymous cloak to legitimize their aggressive actions on social media, the next step is to "*identify what their weakness are and what triggers them - deplatforming or provocation?*" If the misinformers are unresponsive spreaders, then "*exposing, doxing, and putting their real faces through OSINT*" is seen as justified, not just on mainstream social media but also alt-platforms, forums and everywhere on Internet. If they itch for a provocation, then "*orchestrated saturation*" might work better with "*shitposts, absurd trolling, and ridiculing memes*" in the view of our participants. Here, it is vitally important to *a priori* distance from a "*political whataboutery*" and avoid "*coming across as censorship, disagreement, canceling that only could cause argument or dismissal.*"

Some of the hacktivists were on the opinion that "*doxing is not hacking anymore per se because you can get stuff with a credit card and documents could be easily faked nowadays.*" One possible tactic, proposed by one of them, was to "*find exploits, vulnerabilities in their platforms and step-by-step expose misinformers' amateurish way of doing trolling, using bots, and feeding think tanks to get a credibility behind their propaganda.*" Another tactic was "*doxing for the purpose of having advertisers pull from supporting known misinformer influencers, like for example in the case of* [redacted]." Proposing a hybrid style of hacktivist tactics, one participant suggested "*a latent, yet coordinated psychological warfare where psychologists rip apart these people, conduct serious OSINT to find incriminating leaks on them, and even pay for billboards and radio ads to publicly shame them.*" Along these lines, another participant even suggested leveraging all available tactics, "*targeting them with a social engineering attack and compromise a piece of their core infrastructure, be that their servers, Internet access, or bot credentials.*"

7.2 Misinformation Literacy

Hacktivists in our sample echo the sentiment regarding social media users' susceptibility to false information found in scientific literature: laziness to check facts [93], resistance to corrections [59], allegiance [125], and ignorance [19]. As people that resort to action, hacktivists feel the obligation to propose ways to address this susceptibility. In the view of one participant, "*misinformation needs to be seen as something everyone is being watched for, and not just one group of people on the left or the right.*" A "misinformation social contract" [142] necessitates interventions such as "*a critical thinking curricula in schools,*" "*teaching hacking operational security skills as social responsibility and rise to action,*" and "*forcing professional communication norms on platforms.*"

As our participants have little direct control over these interventions, they frequently proposed the development of "*truth-spreading bots for a 'standoff' with misinformation-spreading bots*" as something that could complement the practice of leaks, doxing, and exposure. They recognized that these "truth-spreading bots" must help ordinary users to better find facts, as information literacy is the single most effective tool in dispelling falsehoods [57]. Hacktivists reiterate that platforms do have to let "*misinformation to float on social media and make bots visible, so they gets overwhelmed with factual information*" in order to demonstrate to ordinary users how to do it themselves.

Regardless of whether these stances are realistic or not, the participants in our sample believed that the current approach to improving misinformation literacy is ineffective because

it does not signal an "*unbiased attitude*" to the social media users in the wrong. Instead of an educational and respectable tone, "*rather a 'cancel culture' infused or a 'your opinion is wrong' tone*" plagues any attempt to help people to navigate and locate factual information. Rejection of misinformation, as a result of misinformation literacy, must come as an agreement that "*scientific facts do not have political properties, even if the social media platforms inherently do.*"

7.3 Misinformation Hacktivism

The participants in our sample acknowledge that orchestrated *misinformation hacktivism*, barring individual instances of operations against misinformers, is largely absent from social media. For the hacktivists to assume misinformation as a worthy cause for action, the conflict between the past "hacking for political causes" and [60] future "hacking against using falsehoods in furthering political causes" [24] must be resolved. Though this conflict is complex and evolving, several of the participants worried that it could nevertheless create a "*division between the hacktivists on political lines.*"

As a relative threat to the misinformation activism, one participant mentioned the hijacking of the hacktivists image for self-promotion, e.g. "*some like to portrait themselves as woke gods of the web with zero fuck-ups*." Another threat is the temptation of using misinformation against misinformation, as in the #OpJane campaign. While this strategy is true to the "fight-fire-with-fire" approach, it might backfire in circumstances where abiding to the hacktivist ethic comes secondary to expressing social and political angst on social media [83]. On top of this, one could argue that this conflict *per se* might be hard to resolve in the case of external propaganda, because even if the hacktivists are "hacking for the homeland," they are nonetheless doing it on political terms [28].

8 Discussion

8.1 Implications

The new brand of misinformation, our findings show, draws the attention of the hacktivists, who find the hijacking of discourse for political and propagandistic purposes reprehensible. The "fight-fire-with-fire" response – leaks, doxing, and deplatforming – though individually employed by some of the participants in our sample, has yet to be orchestrated and tested against serious disinformation outfits that, unfortunately, are still prominent on social media [50]. Early evidence from the Ukrainian IT Army's work against Russian wartime propaganda suggests that these new orchestration tactics bring a degree of success [25].

The hacktivists' resolve to go after the misinformers would certainly have implications for the content/user moderation on social media, user participation, and the future of Internet activism overall. Moderating users and content on social media was, and remains, the default response of mainstream platforms to political and public health misinformation [112]. On the other hand, alternative platforms like Gab, Gettr, and Parler, which are seen as breeding grounds for the misinformation [139], have not and currently do not employ the same content and user moderation [111]. While content/user moderation incites a migration from mainstream platforms to the alt-platforms [139], it remains to be seen whether deplatforming will have the same effect. Mainstream social media has had a mixed response to leaks and doxing in the past (e.g. allowing WikiLeaks [118] but barring the Hunter Biden laptop leaks [30]), which adds uncertainty to if and how the hacktivists' "fight-fire-with-fire" approach will be allowed, moderated, or perhaps even forced to migrate entirely outside of the social media space.

Trolling and memes might still maintain popularity amongst misinformers, but, the latest modes of social media participation like short videos on TikTok open new "fronts" for both the misinformers and the hacktivists. TikTok has increasingly been tested as the next "battlefield" of alternative narratives with evidence of health and abortion misinformation [9, 116] and individual engagement by at least one of participants. Recalling that hacktivists' #OpJane was waged in response to the abortion ban laws in Texas and called for "misinformation-against-(mis)information" [41], it is yet to be seen how leaks, doxing, and deplatforming will interact with meme-ified videos and trolling. The company behind TikTok claims it does moderate health and abortion misinformation [129], but evidence shows that this is lax and largely ineffective [16], adding an additional incentive for the adoption of this platform by disinformation campaigns.

TikTok is also poised to become the next platform for Internet activism where the hashtag activism. Here the hashtag activism is combined with videos expanding the developing news narratives, such as the coverage of the Black Lives Matter movement and the Capitol riot [74]. TikTok presents content not just from viral hashtags but also their variations (e.g. #abotion but also #abôrtion [116]) so the threat of hashtag hijacking, co-opting, and counter hash tagging will inevitably materialize here too. This particular affordance will likely facilitate the further weaponization of deepfakes in the near future, as they have already appeared on TikTok in misinformation videos about the COVID-19 pandemic [110]. All of these developments would certainly necessitate a dynamic adaptation in the way doxing, leaking, and deplatforming are performed in order to not just avoid the disintegration of Internet activism and hacktivism, but prevent another paucity in action like the one which brought state-sponsored misinformation *en masse* on social media in the first place [44].

8.2 Ethical Considerations

The purpose of our study was not to generalize to a population, but rather to explore the contemporary hacktivists' relation-

ship with misinformation in depth. To protect individual's privacy, and to avoid speculations, we omitted the names and some of the procedures mentioned during the interviews. We are careful with our study not to infringe upon the hacktivist's sensibilities nor to cause any retaliation with our findings. Though our analysis and interpretation of the findings is positioned on impartiality, the overall study suggests the need for a stronger ethical contextualization of the techniques expressed by the hacktivists, the harms done by this community in the past, and the risk of future harms (intentional or accidental) to individuals and their close ones.

Certain hacktivists have used the techniques discussed in this study to harass, dox and cause harm to intended targets and innocent bystanders alike. For example, the GamerGate scandal emerged as a result of doxing and harassment targeting female gamers – including rape and death threats on a daily basis – by hacktivists that perceived feminism as a threat to traditional values of video games [1]. Similarly, proponents of the #metoo movement have been targets of abusive comments, doxing, and trolling [82]. Even academics and journalists have experienced targeted harassment regardless of the impartiality in their inquiry and "good faith" reporting on the activities associated with the hacktivist communities [32].

These instances clearly contradict Levy's postulate that *computers can change your life for the better* [67] and also undermine the common hacktivist value of defending human rights from any oppression, especially in a sociopolitical, right-of-center context [76]. The ethical justification behind traditional hacktivism as civil disobedience is predicated, and still is, on the premise that "no damage is done to persons" during any hacktivist operation or action [127]. The above actions clearly violate this principle and, as such, lose credibility from a civil disobedience perspective (it is worth pointing out that participants in our study explicitly called for a code of conduct where "*no leak, doxing, or exposure action should cause anyone else harm (physical, reputation, mental)*" towards addressing this issue).

Ethics violations by a hacktivist calls into question the hacktivist's credibility to speak on misinformation, and entails a degree of wariness for future consideration of the proposed anti-misinformation approach. The fact that all the participants in our study, to the best of our knowledge, had no involvement in the aforementioned violations or other unethical activities, cannot alone verify the credibility of the results. Equally, the claims that several participants provided in taking actions *against* sexual harassment and cyberstalking towards women cannot be seen as a vote of confidence that applies to everyone in the sample or the hacktivist community. Our participants or other hacktivists do not always get things right and can be assumed to have the expertise to cover their tracks online [121]. Therefore, the findings reported in our paper do not grant any exemption nor condone engaging in morally dubious or illegal acts.

Our results alone serve neither as an approval nor a call for any hacktivist action. We make no claims to be able to identify and agree with hacktivists on all the "bad actors" in the misinformation space, nor even on a reliable definition of what content constitutes "misinformation" itself [13]. We maintain the caveat that any action – misinformation hacktivism or otherwise – must be morally justified separately. We do, however, identify with the many of the ideas and suggestions posed by the hacktivists in our study – particularly when in conformity with the guidelines put forth in Levy's hacker ethos [67] and when they are in support of a democratic vision of the Internet [45]. We do also support the idea of "fighting fire with fire" identified in our findings inasmuch as it seeks to address the power mismatch that has arisen in the context of regulation, lax policy, and perverse incentive structures on social media platforms – factors which have contributed to the current state of affairs where misinformation is a prominent component of everyday discourse [109]. Again, this is not to legitimize hacktivist actions across the board nor to herald them as the sole defenders but simply to highlight the pressing need for examples of organized resistance against misinformation and well-resourced troll farms online.

8.3 Limitations

Our research was limited in its scope to US-based hacktivists, so we exercise caution not to generalize the results across the entire Internet activist community. Hacktivist operations are often at the center of debates regarding the dimensions of civil disobedience, political participation, legality, and the ethical use of Internet technologies [109]. Our results seek neither to approve or disapprove of these operations, but rather share in-depth accounts of the perspective of this unique and engaged Internet minority. Even with such a relatively small sample, it is clear that hacktivism encompasses a wide variety of perspectives. That being said, we acknowledge the limitations of our sampling and that the individuals in our sample present a limited subset of views and experiences.

We are aware that our results only reflect the current, limited understanding of misinformation informed by the forms currently prominent on social media. Therefore, we are careful to avoid any predictive use of our results in future misinformation campaigns. We also do not know if, when or how the activists in our sample have used the proposed counter-misinformation tools, tactics, and procedures. Our results do not do not offer a blanket justification for the frivolous use of them across any online space. We note that this study reported on the evolving experience of dealing with misinformation by hacktivists and might miss some important aspects of meting out truth on social media. We advise caution in this, as we see our work as a synergistic line of scientific inquiry that addresses an important gap in voicing the opinions of those that actually introduced the means for mass producing of misinformation online in the first place.

8.4 Future Work

Our future research will continue to to trace out the ways in which the hacktivist community engages with misinformation. We plan to expand our work beyond the US and working with hacktivists worldwide, as misinformation influences geopolitical affairs across the globe. We are set to further explore the intersection of and interactions between hashtag activism and hacktivism targeting online misinformation, as synergy between the two have emerged, such as in the case of the #NAFO campaign on Twitter. Here, we would devote much attention to the new misinformation "battlefield" of short-form video platforms such as TikTok. It would be useful to study the emergent circumstances in which misinformation hacktivism mobilizes and empowers ordinary users to join future "Troll [target] Day" operations and and to catalog their experiences with such participation. Of equal importance would be to further study the use of "misinformation-against-(mis)information" as in the case of #OpJane to learn both the useful and harmful aspects of this approach.

9 Conclusion

Reflecting the communitarian ideals of free information and disobedience to authority, the hacktivists in our study showed a determination for a radical response against the reprehensible act of spreading falsehoods on social media. As misinformation is consequential to the trolling and memes of the early days of hacktivism, it is reassuring to observe that the contemporary hacktivists are outwardly against such a nefarious appropriation of their aesthetics. It is encouraging to reveal that hacktivists also advocate for general misinformation literacy as a strategic asset against an undemocratic Internet. These findings, we hope, will empower ordinary users in counteracting misinformation towards the vision of a democratic Internet.

References

[1] Sarah A. Aghazadeh, Alison Burns, Jun Chu, Hazel Feigenblatt, Elizabeth Laribee, Lucy Maynard, Amy L. M. Meyers, Jessica L. O'Brien, and Leah Rufus. *GamerGate: A Case Study in Online Harassment*, pages 179–207. Springer International Publishing, Cham, 2018.

[2] Shiza Ali, Mohammad Hammas Saeed, Esraa Aldreabi, Jeremy Blackburn, Emiliano De Cristofaro, Savvas Zannettou, and Gianluca Stringhini. Understanding the effect of deplatforming on social networks. In *13th ACM Web Science Conference 2021*, WebSci '21, page 187–195, New York, NY, USA, 2021. Association for Computing Machinery.

[3] Jennifer Allen, Cameron Martel, and David G Rand. Birds of a feather don't fact-check each other: Partisanship and the evaluation of news in twitter's birdwatch crowdsourced fact-checking program. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery.

[4] Omar Alonso, Vasileios Kandylas, Serge-Eric Tremblay, Jake M Hofman, and Siddhartha Sen. What's happening and what happened: Searching the social web. In *Proceedings of the 2017 ACM on Web Science Conference*, pages 191–200, 2017.

[5] Janna Anderson and Lee Rainie. The future of truth and misinformation online. 2017. https://www.pewresearch.org/internet/2017/10/19/the-future-of-truth-and-misinformation-online/.

[6] Anonymous. Operation jane initiated. we're totally going to mess with texas. #anonymous, 2021. https://twitter.com/YourAnonNews/status/1433926829396668429.

[7] Ahmer Arif, Leo Graiden Stewart, and Kate Starbird. Acting the part: Examining information operations within# blacklivesmatter discourse. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–27, 2018.

[8] JP Armstrong. Twitter as a channel for frame diffusion? hashtag activism and the virality of# heterosexualprideday. *Rise of the Far Right: Technologies of Recruitment and Mobilization*, page 87, 2021.

[9] Corey H. Basch, Zoe Meleo-Erwin, Joseph Fera, Christie Jaime, and Charles E. Basch. A global pandemic in the time of viral memes: Covid-19 vaccine misinformation and disinformation on tiktok. *Human Vaccines & Immunotherapeutics*, 17(8):2373–2377, 2021.

[10] Ross W. Bellaby. An ethical framework for hacking operations. *Ethical Theory and Moral Practice*, 24(1):231–255, 2021.

[11] Yochai Benkler, Robert Faris, and Hal Roberts. *Network propaganda: Manipulation, disinformation, and radicalization in American politics*. Oxford University Press, 2018.

[12] Davide Beraldo. Unfolding #anonymous on twitter: The networks behind the mask. *First Monday*, 27(1), 2023/01/20 2022. https://firstmonday.org/ojs/index.php/fm/article/view/11723.

[13] Sven Bernecker, Amy K Flowerree, and Thomas Grundmann. *The epistemology of fake news*. Oxford University Press, 2021.

[14] Jonathan Bishop. Trolling for the lulz?: using media theory to understand transgressive humor and other internet trolling in online communities. In *Transforming politics and policy in the digital age*, pages 155–172. IGI Global, 2014.

[15] Amanda S. Bradshaw. #doctorsspeakup: Exploration of hashtag hijacking by anti-vaccine advocates and the influence of scientific counterpublics on twitter. *Health Communication*, 0(0):1–11, 2022.

[16] Jack Brewster, Lorenzo Arvanitis, Valerie Pavilonis, and Macrina Wang. Beware the 'new google:' tiktok's search engine pumps toxic misinformation to its young users, 2022. https://www.newsguardtech.com/misinformation-monitor/september-2022/.

[17] David Bromell. *Deplatforming and Democratic Legitimacy*, pages 81–109. Springer International Publishing, Champaigne, 2022.

[18] Victoria Carty. *Social Movements and New Technology*. Taylor and Francis, 2018.

[19] Jin-Hee Cho, Scott Rager, John O'Donovan, Sibel Adali, and Benjamin D. Horne. Uncertainty-based false information propagation in social networks. *Trans. Soc. Comput.*, 2(2), jun 2019.

[20] Miyoung Chong. Discovering fake news embedded in the opposing hashtag activism networks on twitter: #gunreformnow vs. #nra. *Open Information Science*, 3(1):137–153, 2019.

[21] Matteo Cinelli, Walter Quattrociocchi, Alessandro Galeazzi, Carlo Michele Valensise, Emanuele Brugnoli, Ana Lucia Schmidt, Paola Zola, Fabiana Zollo, and Antonio Scala. The covid-19 social media infodemic. *Scientific reports*, 10(1):1–10, 2020.

[22] E Gabriella Coleman. Logics and legacy of anonymous. *Second International Handbook of Internet Research*, pages 145–166, 2020.

[23] Gabriella Coleman. *Hacker, hoaxer, whistleblower, spy: The many faces of Anonymous*. Verso books, 2014.

[24] Christopher T Conner and Nicholas MacMurray. The perfect storm: A subcultural analysis of the q-anon movement. *Critical Sociology*, 48(6):1049–1071, 2022.

[25] Ellen Cornelius. Anonymous Hacktivism: Flying the Flag of Feminist Ethics for the Ukraine IT Army. 2022.

[26] John Corner. Fake news, post-truth and media–political change, 2017.

[27] Brian Creech. Fake news and the discursive construction of technology companies' social power. *Media, Culture & Society*, 42(6):952–968, 2020.

[28] Michael Dahan. Hacking for the homeland: Patriotic hackers versus hacktivists. In *Proceedings of the 8th International Conference on Information Warfare and Security (Iciw-2013)*, pages 51–57, 2013.

[29] Philipp Darius and Fabian Stephany. How the far-right polarises twitter: 'hashjacking' as a disinformation strategy in times of covid-19. In Rosa Maria Benito, Chantal Cherifi, Hocine Cherifi, Esteban Moro, Luis M. Rocha, and Marta Sales-Pardo, editors, *Complex Networks & Their Applications X*, pages 100–111, Cham, 2022. Springer International Publishing.

[30] Glenn Diesen. *Conclusion: Anti-Russian Propaganda of a West in Relative Decline*, pages 255–258. Springer Nature Singapore, Singapore, 2022.

[31] Renee DiResta, Kris Shaffer, Becky Ruppel, David Sullivan, Robert Matney, Ryan Fox, Jonathan Albright, and Ben Johnson. The tactics & tropes of the internet research agency. 2019.

[32] Periwinkle Doerfler, Andrea Forte, Emiliano De Cristofaro, Gianluca Stringhini, Jeremy Blackburn, and Damon McCoy. "i'm a professor, which isn't usually a dangerous job": Internet-facilitated harassment and its impact on researchers. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–32, 2021.

[33] Ullrich K. H. Ecker, Stephan Lewandowsky, John Cook, Philipp Schmid, Lisa K. Fazio, Nadia Brashier, Panayiota Kendeou, Emily K. Vraga, and Michelle A. Amazeen. The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, 1(1):13–29, 2022.

[34] Abbas Ehsanfar and Mo Mansouri. Incentivizing the dissemination of truth versus fake news in social networks. In *2017 12th System of Systems Engineering Conference (SoSE)*, pages 1–6, 2017.

[35] Luca Follis and Adam Fish. *3 When to Hack*, pages 73–111. 2020.

[36] Deen Freelon, Alice Marwick, and Daniel Kreiss. False equivalencies: Online activism from left to right. *Science*, 369(6508):1197–1201, 2020.

[37] Deen Freelon, Charlton D McIlwain, and Meredith Clark. Beyond the hashtags:# ferguson,# blacklivesmatter, and the online struggle for offline justice. *Center for Media & Social Impact, American University, Forthcoming*, 2016.

[38] Gordon Gauchat. Politicization of science in the public sphere: A study of public trust in the united states, 1974 to 2010. *American sociological review*, 77(2):167–187, 2012.

[39] Jordana J. George and Dorothy E. Leidner. From clicktivism to hacktivism: Understanding digital activism. *Information and Organization*, 29(3):100249, 2019.

[40] Paolo Gerbaudo. From cyber-autonomism to cyber-populism: An ideological history of digital activism. *tripleC: Communication, Capitalism & Critique*, 15(2):477–489, May 2017.

[41] Claire Goforth. 'anonymous' hackers have a message for texas abortion 'snitch' sites: We're coming for you, 2021. https://www.dailydot.com/debug/anonymous-hactivists-texas-abortion-ban-operation-jane/.

[42] Atilla Hallsby. Psychoanalysis against wikileaks: resisting the demand for transparency. *Review of Communication*, 20(1):69–86, 2020.

[43] Max Halupka. Clicktivism: A systematic heuristic. *Policy & Internet*, 6(2):115–132, 2014.

[44] Jason Hannan. Trolling ourselves to death? social media and post-truth politics. *European Journal of Communication*, 33(2):214–226, 2018.

[45] Masayuki Hatta. Cowboys and the eternal september transfiguration of hacker aesthetics. *Annals of Business Administrative Science*, page 0210923a, 2021.

[46] Nolan Higdon, Emil Marmol, and Mickey Huff. Returning to neoliberal normalcy: Analysis of legacy news media's coverage of the biden presidency's first hundred days. In *The Future of the Presidency, Journalism, and Democracy*, pages 255–273. Routledge, 2022.

[47] Matthew Hindman and Vlad Barash. Disinformation, 'fake news' and influence campaigns on twitter. 2018.

[48] Philip N Howard, Bharath Ganesh, Dimitra Liotsiou, John Kelly, and Camille François. The IRA, social media and political polarization in the United States, 2012-2018. 2019.

[49] Laura Illia. Passage to cyberactivism: how dynamics of activism change. *Journal of public affairs.*, 3(4), 2003-11.

[50] Jane Im, Eshwar Chandrasekharan, Jackson Sargent, Paige Lighthammer, Taylor Denby, Ankit Bhargava, Libby Hemphill, David Jurgens, and Eric Gilbert. Still out there: Modeling and identifying russian troll accounts on twitter. In *12th ACM Conference on Web Science*, WebSci '20, page 1–10, New York, NY, USA, 2020. Association for Computing Machinery.

[51] Leanna Ireland. We are all (not) anonymous: Individual- and country-level correlates of support for and opposition to hacktivism. *New Media & Society*, 0(0):14614448221122252, 0.

[52] Peter Jachim, Filipo Sharevski, and Emma Pieroni. Trollhunter2020: Real-time detection of trolling narratives on twitter during the 2020 us elections. In *Proceedings of the 2021 ACM workshop on security and privacy analytics*, pages 55–65, 2021.

[53] Peter Jachim, Filipo Sharevski, and Paige Treebridge. Trollhunter [evader]: Automated detection [evasion] of twitter trolls during the covid-19 pandemic. In *New Security Paradigms Workshop 2020*, NSPW '20, page 59–75, New York, NY, USA, 2021. Association for Computing Machinery.

[54] S.J. Jackson, M. Bailey, B.F. Welles, and G. Lauren. *#HashtagActivism: Networks of Race and Gender Justice*. MIT Press, 2020.

[55] Keenan Jones, Jason R. C. Nurse, and Shujun Li. Behind the mask: A computational study of anonymous' presence on twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1):327–338, May 2020.

[56] Keenan Jones, Jason R.C. Nurse, and Shujun Li. Out of the shadows: Analyzing anonymous' twitter resurgence during the 2020 black lives matter protests. *Proceedings of the International AAAI Conference on Web and Social Media*, 16(1):417–428, May 2022.

[57] S Mo Jones-Jang, Tara Mortensen, and Jingjing Liu. Does media literacy help identification of fake news? information literacy helps, but other literacies don't. *American behavioral scientist*, 65(2):371–388, 2021.

[58] Andreas Jungherr, Gonzalo Rivero, and Daniel Gayo-Avello. *Retooling Politics: How Digital Media Are Shaping Democracy*. Cambridge University Press, 2020.

[59] Ben Kaiser, Jerry Wei, Eli Lucherini, Kevin Lee, J Nathan Matias, and Jonathan R Mayer. Adapting security warnings to counter online disinformation. In *USENIX Security Symposium*, pages 1163–1180, 2021.

[60] Vasileios Karagiannopoulos. *A Short History of Hacktivism: Its Past and Present and What Can We Learn from It*, pages 63–86. Springer International Publishing, Cham, 2021.

[61] Athina Karatzogianni. *Firebrand waves of digital activism 1994-2014: The rise and spread of hacktivism and cyberconflict*. Springer, 2015.

[62] Matthew D Kearney, Shawn C Chiang, and Philip M Massey. The twitter origins and evolution of the covid-19 "plandemic" conspiracy theory. *Harvard Kennedy School Misinformation Review*, 1(3), 2020.

[63] Vance D. Keyes and Latocia Keyes. Dynamics of an american countermovement: Blue lives matter. *Sociology Compass*, 16(9):e13024, 2022.

[64] Allison Klempka and Arielle Stimson. Anonymous communication on the internet and trolling. *Concordia Journal of Communication Research*, 1(1):2, 2014.

[65] Deepak Kumar, Patrick Gage Kelley, Sunny Consolvo, Joshua Mason, Elie Bursztein, Zakir Durumeric, Kurt Thomas, and Michael Bailey. Designing toxic content classification for a diversity of perspectives. In *SOUPS@ USENIX Security Symposium*, pages 299–318, 2021.

[66] Micah Lee. *Hack of 251 Law Enforcement Webiste Exposes Personal Data of 700,000 Cops*. 2020. https://theintercept.com/2020/07/15/blueleaks-anonymous-ddos-law-enforcement-hack/.

[67] Steven Levy. *Hackers: Heroes of the Computer Revolution - 25th Anniversary Edition*. O'Reilly Media, Inc., 1st edition, 2010.

[68] Stephan Lewandowsky, John Cook, Ullrich Ecker, Dolores Albarracin, Michelle Amazeen, Panayiota Kendou, Doug Lombardi, E Newman, Gordon Pennycook, Ethan Porter, et al. *The Debunking Handbook 2020*. 2020.

[69] Stephan Lewandowsky, Ullrich K.H. Ecker, and John Cook. Beyond misinformation: Understanding and coping with the "post-truth" era. *Journal of Applied Research in Memory and Cognition*, 6(4):353–369, 2017.

[70] Stephan Lewandowsky and Sander van der Linden. Countering misinformation and fake news through inoculation and prebunking. *European Review of Social Psychology*, 32(2):348–384, 2021.

[71] Simon Lindgren. Movement mobilization in the age of hashtag activism: Examining the challenge of noise, hate, and disengagement in the #metoo campaign. *Policy & Internet*, 11(4):418–438, 2019.

[72] Alexander J Lindvall. Political hacktivism: doxing & the first amendment. *Creighton L. Rev.*, 53:1, 2019.

[73] Darren L. Linvill and Patrick L. Warren. Troll factories: Manufacturing specialized disinformation on twitter. *Political Communication*, 37(4):447–467, 2020.

[74] Ioana Literat, Lillian Boxman-Shabtai, and Neta Kligler-Vilenchik. Protesting the protest paradigm: Tiktok as a space for media criticism. *The International Journal of Press/Politics*, 0(0):19401612221117481, 0.

[75] Clare Llewellyn, Laura Cram, Adrian Favero, and Robin L. Hill. Russian troll hunting in a brexit twitter archive. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*, JCDL '18, page 361–362, New York, NY, USA, 2018. Association for Computing Machinery.

[76] Mark Manion and Abby Goodrum. Terrorism or civil disobedience: toward a hacktivist ethic. *Acm Sigcas Computers and Society*, 30(2):14–19, 2000.

[77] Martha McCaughey and Michael D Ayers. *Cyberactivism: Online activism in theory and practice*. Psychology Press, 2003.

[78] Ty McCormick. Anthropology of an idea hacktivism. *Foreign Policy*, (200):24–25, May/Jun 2013.

[79] Ally McCrow-Young and Mette Mortensen. Countering spectacles of fear: Anonymous' meme 'war' against isis. *European Journal of Cultural Studies*, 24(4):832–849, 2021.

[80] Virginia McGovern and Francis Fortin. The anonymous collective: Operations and gender differences. *Women & Criminal Justice*, 30(2):91–105, 2020.

[81] Lee McIntyre. *Post-truth*. MIT Press, 2018.

[82] Kaitlynn Mendes, Jessica Ringrose, and Jessalynn Keller. # metoo and the promise and pitfalls of challenging rape culture through digital feminist activism. *European Journal of Women's Studies*, 25(2):236–246, 2018.

[83] Paul Mihailidis and Samantha Viotty. Spreadable spectacle in digital culture: Civic expression, fake news, and the role of media literacies in "post-fact" society. *American behavioral scientist*, 61(4):441–454, 2017.

[84] Stefania Milan. *Social movements and their technologies: Wiring social change*. Springer, 2013.

[85] Stefania Milan. Hacktivism as a radical media practice. In *The Routledge companion to alternative and community media*, pages 550–560. Routledge, 2015.

[86] Ryan M Milner. *The world made meme: Public conversations and participatory media*. MIT Press, 2018.

[87] Rachel E. Moran and Stephen Prochaska. Misinformation or activism?: analyzing networked moral panic through an exploration of #savethechildren. *Information, Communication & Society*, 0(0):1–21, 2022.

[88] Mette Mortensen and Christina Neumayer. The playful politics of memes. *Information, Communication & Society*, 24(16):2367–2377, 2021.

[89] Eni Mustafaraj and Panagiotis Takis Metaxas. The fake news spreading plague: Was it preventable? In *Proceedings of the 2017 ACM on Web Science Conference*, WebSci '17, page 235–239, New York, NY, USA, 2017. Association for Computing Machinery.

[90] Mahdi M. Najafabadi and Robert J. Domanski. Hacktivism and distributed hashtag spoiling on twitter: Tales of the #irantalks. *First Monday*, 23(4), Apr. 2018. https://firstmonday.org/ojs/index.php/fm/article/view/8378.

[91] Asaf Nissenbaum and Limor Shifman. Internet memes as contested cultural capital: The case of 4chan's/b/board. *New media & society*, 19(4):483–501, 2017.

[92] Taylor Owen. *Disruptive power: The crisis of the state in the digital age*. Oxford Studies in Digital Politics, 2015.

[93] Gordon Pennycook and David G. Rand. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, 188:39–50, 2019. The Cognitive Science of Political Thought.

[94] Gordon Pennycook and David G. Rand. The psychology of fake news. *Trends in Cognitive Sciences*, 25(5):388–402, 2021.

[95] W. Phillips and R.M. Milner. *The Ambivalent Internet: Mischief, Oddity, and Antagonism Online*. Polity Press, 2017.

[96] Whitney Phillips. The house that fox built: Anonymous, spectacle, and cycles of amplification. *Television & New Media*, 14(6):494–509, 2013.

[97] Man pui Sally Chan, Christopher R. Jones, Kathleen Hall Jamieson, and Dolores Albarracín. Debunking: A meta-analysis of the psychological efficacy of messages countering misinformation. *Psychological Science*, 28(11):1531–1546, 2017.

[98] David Redding, Jian Ang, and Suman Bhunia. A case study of massive api scrapping: Parler data breach after the capitol riot. In *2022 7th International Conference on Smart and Sustainable Technologies (SpliTech)*, pages 1–7, 2022.

[99] Eugenia Ha Rim Rho and Melissa Mazmanian. Political hashtags & the lost art of democratic discourse. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–13, New York, NY, USA, 2020. Association for Computing Machinery.

[100] Daniel R´ochert, Gautam Kishore Shahi, German Neubaum, Björn Ross, and Stefan Stieglitz. The networked context of covid-19 misinformation: Informational homogeneity on youtube at the beginning of the pandemic. *Online Social Networks and Media*, 26:100164, 2021.

[101] Mark Rolfe. *Hacker: Creating the Narrative of the Digital Robin Hood*, pages 135–164. Springer Singapore, 2016.

[102] Marco Romagna. *Hacktivism: Conceptualization, Techniques, and Historical View*, pages 743–769. Springer International Publishing, Cham, 2020.

[103] Dana Rotman, Sarah Vieweg, Sarita Yardi, Ed Chi, Jenny Preece, Ben Shneiderman, Peter Pirolli, and Tom Glaisyer. From slacktivism to activism: Participatory culture in the age of social media. In *CHI '11 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '11, page 819–822, New York, NY, USA, 2011. Association for Computing Machinery.

[104] Rodrigo Sandoval-Almazan and J. Ramon Gil-Garcia. Towards cyberactivism 2.0? understanding the use of social media and other information technologies for political activism and social movements. *Government Information Quarterly*, 31(3):365–378, 2014.

[105] Madelyn R Sanfilippo, Shengnan Yang, and Pnina Fichman. Managing online trolling: From deviant to social and political trolls. In *50th Annual Hawaii International Conference on System Sciences, HICSS 2017*, pages 1802–1811. IEEE Computer Society, 2017.

[106] Saiph Savage, Andres Monroy-Hernandez, and Tobias Höllerer. Botivist: Calling volunteers to action using online bots. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, CSCW '16, page 813–822, New York, NY, USA, 2016. Association for Computing Machinery.

[107] Mark Scott. The shit-posting, twitter-trolling, dog-deploying social media army taking on putin one meme at a time, 2022.

[108] Dimitrios Serpanos and Theodoros Komninos. The cyberwarfare in ukraine. *Computer*, 55(7):88–91, 2022.

[109] Philip Serracino-Inglott. Is it ok to be an anonymous? *Ethics & Global Politics*, 6(4):22527, 2013.

[110] Lanyu Shang, Ziyi Kou, Yang Zhang, and Dong Wang. A multimodal misinformation detector for covid-19 short videos on tiktok. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 899–908, 2021.

[111] Filipo Sharevski, Amy Devine, Peter Jachim, and Emma Pieroni. "Gettr-ing" User Insights from the Social Network Gettr, 2022. https://truthandtrustonline.com/wp-content/uploads/2022/10/TTO_2022_proceedings.pdf.

[112] Filipo Sharevski, Amy Devine, Peter Jachim, and Emma Pieroni. Meaningful context, a red flag, or both? preferences for enhanced misinformation warnings among us twitter users. In *Proceedings of the 2022 European Symposium on Usable Security*, EuroUSEC '22, page 189–201, New York, NY, USA, 2022. Association for Computing Machinery. https://doi.org/10.1145/3549015.3555671.

[113] Filipo Sharevski, Amy Devine, Emma Pieroni, and Peter Jachim. Folk models of misinformation on social media. In *Network and distributed system security symposium*, 2023.

[114] Filipo Sharevski, Alice Huff, Peter Jachim, and Emma Pieroni. (mis)perceptions and engagement on twitter: Covid-19 vaccine rumors on efficacy and mass immunization effort. *International Journal of Information Management Data Insights*, 2(1):100059, 2022.

[115] Filipo Sharevski, Peter Jachim, Emma Pieroni, and Nate Jachim. Voxpop: An experimental social media platform for calibrated (mis)information discourse. In *New Security Paradigms Workshop*, NSPW '21, page 88–107, New York, NY, USA, 2021. Association for Computing Machinery.

[116] Filipo Sharevski, Jennifer Vander Loop, Peter Jachim, Amy Devine, and Emma Pieroni. Abortion misinformation on tiktok: Rampant content, lax moderation, and vivid user experiences. *arXiv preprint arXiv:2301.05128*, 2023.

[117] Filipo Sharevski, Paige Treebridge, Peter Jachim, Audrey Li, Adam Babin, and Jessica Westbrook. Socially engineering a polarizing discourse on facebook through malware-induced misperception. *International Journal of Human–Computer Interaction*, 38(17):1621–1637, 2022.

[118] Micah L Sifry. *WikiLeaks and the Age of Transparency*. OR Books, 2011.

[119] Ellen Simpson. Integrated & alone: The use of hashtags in twitter social activism. In *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '18, page 237–240, New York, NY, USA, 2018. Association for Computing Machinery.

[120] Edward Snowden. *Permanent record*. Pan Macmillan, 2019.

[121] Tom Sorell. Human Rights and Hacktivism: The Cases of Wikileaks and Anonymous. *Journal of Human Rights Practice*, 7(3):391–410, 09 2015.

[122] Simon Springer, Kean Birch, and Julie MacLeavy. *The handbook of neoliberalism*, volume 12. Routledge New York, 2016.

[123] Kevin F Steinmetz. Hacking and hacktivism. *Shades of Deviance: A Primer on Crime, Deviance and Social Harm*, 19, 2022.

[124] Leo G Stewart, Ahmer Arif, and Kate Starbird. Examining trolls and polarization with a retweet network. In *Proc. ACM WSDM, workshop on misinformation and misbehavior mining on the web*, volume 70, 2018.

[125] Briony Swire-Thompson, Ullrich KH Ecker, Stephan Lewandowsky, and Adam J Berinsky. They might be a liar but they're my liar: Source evaluation and the prevalence of misinformation. *Political psychology*, 41(1):21–34, 2020.

[126] Leonie Maria Tanczer. Hacktivism and the male-only stereotype. *New Media & Society*, 18(8):1599–1615, 2016.

[127] Paul Taylor. Hacktivism: in search of lost ethics? In *Crime and the Internet*, pages 71–85. Routledge, 2003.

[128] William Theisen, Joel Brogan, Pamela Bilo Thomas, Daniel Moreira, Pascal Phoa, Tim Weninger, and Walter Scheirer. Automatic discovery of political meme genres with diverse appearances. *Proceedings of the International AAAI Conference on Web and Social Media*, 15(1):714–726, May 2021.

[129] TikTok. Tiktok safety, 2022. https://www.tiktok.com/safety/en-us/topics/.

[130] Justus Uitermark. Complex contention: analyzing power dynamics within anonymous. *Social Movement Studies*, 16(4):403–417, 2017.

[131] Anthony Vance, David Eargle, Jeffrey L. Jenkins, C. Brock Kirwan, and Bonnie Brinton Anderson. The Fog of Warnings: How Non-essential Notifications Blur with Security Warnings. In *Fifteenth Symposium*

on Usable Privacy and Security (SOUPS 2019), Santa Clara, CA, August 2019. USENIX Association.

[132] Courtland VanDam and Pang-Ning Tan. Detecting hashtag hijacking from twitter. In *Proceedings of the 8th ACM Conference on Web Science*, WebSci '16, page 370–371, New York, NY, USA, 2016. Association for Computing Machinery.

[133] Silvio Waisbord. Truth is what happens to news. *Journalism Studies*, 19(13):1866–1878, 2018.

[134] Jared M Wright, Kaitlin Kelly-Thompson, S Laurel Weldon, Dan Goldwasser, Rachel L Einwohner, Valeria Sinclair-Chapman, and Fernando Tormos-Aponte. Drive-by solidarity: Conceptualizing the temporal relationship between# blacklivesmatter and anonymous's# opkkk. *Contention*, 10(2):25–55, 2022.

[135] Liang Wu, Fred Morstatter, Kathleen M. Carley, and Huan Liu. Misinformation in social media: Definition, manipulation, and detection. *SIGKDD Explor. Newsl.*, 21(2):80–90, nov 2019.

[136] Yiping Xia, Josephine Lukito, Yini Zhang, Chris Wells, Sang Jung Kim, and Chau Tong. Disinformation, performed: self-presentation of a russian ira account on twitter. *Information, Communication & Society*, 22(11):1646–1664, 2019.

[137] Weiai Wayne Xu. Mapping connective actions in the global alt-right and antifa counterpublics. *International Journal of Communication*, 14:22, 2020.

[138] Savvas Zannettou, Barry Bradlyn, Emiliano De Cristofaro, Haewoon Kwak, Michael Sirivianos, Gianluca Stringini, and Jeremy Blackburn. What is gab: A bastion of free speech or an alt-right echo chamber. In *Companion Proceedings of the Web Conference 2018*, pages 1007–1014, 2018.

[139] Savvas Zannettou, Tristan Caulfield, Jeremy Blackburn, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Guillermo Suarez-Tangil. On the origins of memes by means of fringe web communities. In *Proceedings of the Internet Measurement Conference 2018*, IMC '18, page 188–202, New York, NY, USA, 2018. Association for Computing Machinery.

[140] Savvas Zannettou, Tristan Caulfield, Barry Bradlyn, Emiliano De Cristofaro, Gianluca Stringhini, and Jeremy Blackburn. Characterizing the use of images in state-sponsored information warfare operations by russian trolls on twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1):774–785, May 2020.

[141] Savvas Zannettou, Tristan Caulfield, William Setzer, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. Who let the trolls out? towards understanding state-sponsored trolls. In *Proceedings of the 10th ACM Conference on Web Science*, WebSci '19, pages 353–362, New York, NY, USA, 2019. Association for Computing Machinery.

[142] Melissa Zimdars and Kembrew McLeod. *Fake news: understanding media and misinformation in the digital age*. MIT Press, 2020.

Appendix

1. How do you describe your niche, role, activity, or agenda you have online?
2. What brought you to hacking, OSINT, cyber-threat intelligence, and any operations you have taken so far?
3. Have you faced any obstacles, challenges, repercussions because of your activity?
4. Has the obstacles, challenges, repercussions affected your commitment, motivation, and vision of your actions and in what way?
5. What is your take on the increased misinformation proliferation online?
6. Have you ever engaged or considered engaging in utilizing your actions in exposing disinformation campaigns? What was the disinformation about, in what capacity you participated, and what were the outcomes you were attempting to achieve?
7. What do you think the tools, tactics, and procedures undertaken in a hypothetical *misinformation hacktivism* operation might entail?
8. What in your opinion, is the way to continue evolving this work and in what shape and form?
9. Is there anything else that you would like to add or say that is relevant to the questions we have asked so far?
10. If you would like to share some demographic information, please do - we don't require it but it will help us better contextualize your effort and story.

“Stalking is immoral but not illegal": Understanding Security, Cyber Crimes and Threats in Pakistan

Afaq Ashraf
Lahore University of Management Sciences

- Taha
Lahore University of Management Sciences

Nida ul Habib Bajwa
Universität des Saarlandes

Cornelius J. König
Universität des Saarlandes

Mobin Javed
Lahore University of Management Sciences

Maryam Mustafa
Lahore University of Management Sciences

Abstract

We explore the experiences, understandings and perceptions of cyber-threats and crimes amongst young adults in Pakistan, focusing on their mechanisms for protecting themselves, for reporting cyber threats and for managing their digital identities. Relying on data from a qualitative study with 34 participants in combination with a repertory grid analysis with 18 participants, we map users mental models and constructs of cyber crimes and threats, their understanding of digital vulnerabilities, their own personal boundaries and their moral compasses on what constitutes an invasion of privacy of other users in a country where there is little legal legislation governing cyberspace and cyber crimes. Our findings highlight the importance of platform adaptation to accommodate the unique context of countries with limited legal mandates and reporting outlets, the ways in which digital vulnerabilities impact diverse populations, and how security and privacy design can be more inclusive.

1 Introduction

We unpack the experiences, perceptions of cybercrimes and mechanisms for self-protection of young adults in Pakistan, focusing on their social media usage. The pandemic has accelerated the growth of the internet and resulted in a significant increase in internet traffic [14, 23]. This shift has enabled the migration of various activities such as shopping, education, work, and entertainment to online platforms. However, with the rise in internet usage, the number of reported cybercrime incidents has also increased, as noted by the FBI and Interpol in the United Kingdom and the United States [1, 4]. The FBI reported that the number of cybercrime complaints received between January and May 2020 was nearly equivalent to the total number of complaints received in the entire year of 2019 [49]. Similarly, according to the Pakistan Telecommunication Authority (PTA), the number of cybercrime complaints received by the authority has been increasing in recent years. In 2020, the PTA received over 17,000 complaints related to cybercrime, an increase of around 40% compared to the previous year [6]. Despite the increasing number of cybercrime complaints, the conviction rate for cybercriminals in Pakistan remains low [6]. This is due to a number of factors, including a lack of technical expertise among law enforcement agencies, a weak legal framework for combating cybercrime, and a lack of awareness among the general public about how to protect themselves from cybercrime.

Among internet users in Pakistan, young adults constitute a large percentage of the demographic. In 2016, 19% of all internet usage in Pakistan was attributed to individuals aged 15-24 years old [2]. This demographic is considered tech-literate, digitally savvy, and early adopters of online platforms, primarily using social media platforms like Twitter, Instagram, Tiktok and Snapchat. The Digital Rights Foundation, a non-profit operating in Pakistan, reports that approximately 69% of the calls they received reporting cybercrimes were made by individuals within the age range of 18 to 30, with 78% of those calls being made by women, indicating that younger women are disproportionately affected by cybercrimes. There are, however, few studies that explore the relationships young, tech-savvy users in contexts like Pakistan, which have limited legal frameworks governing online spaces and few mechanisms for redressal, have with privacy and security online. Pakistan is a religious, patriarchal cultural context with a strong emphasis on honor and social status, resulting in often extreme consequences of online privacy breaches and harms. This is particularly true for women, as evidenced by high-profile cases such as the honor killing of Qandeel Baloch in Pakistan [27] and the suicide of Vinupriya in India [48]. It is important to unpack *what* a cybercrime constitutes in such a

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.
USENIX Symposium on Usable Privacy and Security (SOUPS) 2023.
August 6–8, 2023, Anaheim, CA, USA

context. What constitutes experienced *harm* for young people and where do platforms fail to account for diverse and complex contexts with varying factors at play, such as patriarchal structures of control and religious connotations?

We gathered qualitative data from 34 interviews and utilized the repertory grid technique (RGT) to collect data from 18 interviews with literate young men and women aged 18 to 23. Our study does not specifically aim to target individuals who have experienced cybercrime. Instead, we focus on literate users at the undergraduate level who have adopted devices and started going online at an early age (some as early as 5 years old). We find their use predominantly centers around social media platforms and much of their experiences and harms are associated with these platforms. Our work makes three key contributions:

1. We unpack what constitutes a cybercrime in this context, laying out the conditions under which online behavior is considered a harm and a crime from the perspective of young people. We also visualize the experienced severity of cybercrime through a gender disaggregated spectrum.

2. We highlight the strategies and behaviors that young users employ to protect themselves on social media platforms and their source of learning for these behaviors.

3. We explore design implications to improve knowledge of privacy mechanisms and increase control over online data sharing among social media users.

2 Related work

The present study expands the existing literature on privacy perceptions by specifically focusing on the perception of *cybercrime* in Pakistan. An understanding of *what* constitutes a cybercrime is complex, and its definition varies depending on the context, user mental models, and perceptions. Prior work in privacy has explored user perceptions and experiences of cybercrimes but often in developed contexts and often with adult populations [35, 51, 53]. Our work aims to fill this gap in knowledge by investigating the perceptions of cybercrime among young adults in the complicated context of Pakistan.

2.1 Privacy Perceptions Across Cultures

Privacy perceptions are influenced by factors such as social norms, individual characteristics, and community dynamics [36]. As a result, individuals in the Global North and Global South have different privacy expectations, beliefs, and behaviours [39]. One study across eight countries reports that Japanese and German participants have higher privacy concerns on their smartphones than those from Australia, Canada, Italy, Netherlands, the UK, and the USA, even though German users rated the sensitivity of their data as lower than that of Italian and Japanese [28]. Phone locks and pins are often used to safeguard data [12, 28, 29]. In contrast, prior work reveals that in South Africa, users are more worried about *who* is viewing their data rather than the data itself or its collection by platforms. These users are unaware of finer privacy features on social media platforms and rely heavily on blocking as the main privacy mechanism [43]. Similarly, Bellman et al. conduct a survey with 534 responses to understand the differences in privacy concerns amongst users across 38 countries. They highlight the importance of cultural values, internet experiences, and desires of political institutes as key factors impacting privacy concerns [11]. Similarly, other work exploring privacy practices of women in Pakistan, India, and Bangladesh reveal five key practices for safeguarding data, including phone locks, app locks, aggregate and entity deletion, private modes, and avoidance [47].

In Pakistan, religion plays a powerful role in shaping privacy attitudes and behaviors where women often negotiate the creation of gendered spaces online as a way of protecting their information from unfamiliar individuals, as prescribed by Islamic teachings [39]. Similarly, Arabic social media users frequently establish private online accounts to uphold *ird*, an Arabic term referring to personal or familial honor, in line with social norms [7]. Muslim women in the USA also reveal how social surveillance, which refers to performing actions out of social obligations within the community, impacted their activities online [8]. Prior work also reveals low-income, low-literate women in Pakistan, India, and Bangladesh associate privacy with *Western values* [47] or with shame [39], often believing privacy is only for people who have done something that they want to hide.

2.2 Cybercrime Experiences and Perceptions

Perceptions of cybercrime vary between the Global North and Global South, with factors such as socioeconomic status, gender, customs, and religion influencing their formation.

An increase in the use of digital spaces and platforms has resulted in a subsequent increase in cybercrimes [15, 30, 37] with online shopping fraud, online banking fraud, cyberbullying like stalking and threatening, malware, and hacking the most prevalent forms of online crimes [42]. In Finland, the 5 most common forms of victimization experienced by people in the age group of 15-74 years old were malware, harassment like defamation and threat of violence, hacking, fraud, and sexual harassment. People with higher internet usage were more impacted by malware attacks [37]. The PEW Research Center conducted a survey in the USA which found that Americans have reported personally experiencing online harassment and view it as a significant issue [21]. Another study by the same research center revealed that 26% of women reported being stalked online, and 25% reported experiencing sexual harassment online [55]. In a survey conducted by Maple et al. with 353 participants, 324 reported experiencing online harassment, with women citing “fear of personal injury" as their top

concern when engaging online, followed by concerns related to their reputation [34].

During the COVID-19 lockdown in the UK, incidents related to "frauds associated with online shopping and auctions, and the hacking of social media and email" saw the largest increases, being the two most common categories of cybercrime in the country [17]. Another study exploring experiences of influencers on TikTok, Facebook, Instagram, Twitter, and YouTube found that at least 95% of creators described facing some form of harassment at least once in their career [52]. The study revealed through a longitudinal study that hate and harassment have grown 4% over the last three years and now affect 48% of people globally. They found that young adults, LGBTQ+ individuals, and individuals who frequently use the internet are more at risk of privacy violations [51].

In contrast, few studies focus on unpacking privacy experiences and behaviours in the Global South. One such study exploring online privacy perceptions and practices in Ghana reveals a lack of understanding of how internet technologies operate with users relying heavily on passwords, and those who augment their security do so with a variety of ad-hoc practices learned through word of mouth [18]. Another study by Sambasivan et al. in India, Pakistan, and Bangladesh, found that a majority of the participants regularly faced online abuse, experiencing three major types: cyberstalking, impersonation, and personal content leakages [46]. Other work reveals Low Socioeconomic Arabs (LSA) experienced black hat hacking, identity theft, shoulder surfing, and defamation. High Socioeconomic Arabs (HSA) experience grey hat hacking, credit card theft, financial fraud, and identity theft. The consequences of the attacks were more severe for LSA like reputational harm while HSA reported little to no consequences from the attacks [45]. According to a 2021 Digital Rights Foundation report, the most commonly reported cyber harassment cases were blackmailing, non-consensual use of information, unsolicited contact, hacked account, financial fraud, fake profile, and defamation. Other threats like impersonation, bullying, hate speech, and cyberstalking were also reported [3].

2.3 Cybercrime among Young Adults

Young people are the most frequent users of the internet and technology, and as such, they are most likely to be exposed to cybercrimes than other demographic groups [13, 56]. This is due to their new financial responsibilities, social independence, and frequent technology usage [13]. While few studies globally have focused on young people those that have report that students often receive messages that threatened, insulted, or harassed them or were pornographic in nature [24, 40]. Prior work also reveals that amongst undergraduate and graduate female students, those who viewed social media as having a negative impact on their lives also reported experiencing more online harassment [54]. Another four-country (Finland, US, UK, Germany) study examining cybercrime victimization among teenagers and young adults, found that online crime victimization was relatively uncommon, with slander and the threat of violence being the most common forms of victimization, and sexual harassment the least common [38]. Crimes like malware, hacking, and phishing were more common among U.S. undergraduate students. The students reported gaining knowledge about cybercrime through prior victims and media sources [13].

Previous studies conducted on technology usage amongst adults in low-literate and low-income areas of Pakistan and other Southeast Asian nations have revealed disparities in technology utilization between men and women and variations in privacy perceptions based on cultural and religious values [39, 46, 47]. This study examines whether young adults with a higher level of education, technological literacy, and economic stability compared to their low-literate and low-income counterparts experience similar challenges with privacy and cybercrime. Furthermore, this study aims to explore the reasons behind these difficulties, if present, despite the higher level of technological literacy in this demographic. Additionally, this study seeks to complement existing literature on privacy and cybercrime, which primarily focuses on the female perspective, by examining the male perspective on these issues.

3 Methodology

This study explores the young adults' mental models of cybercrime in Pakistan. Our main questions were:

- RQ1: What is considered a cybercrime from the perspective of young people, and how do they define the severity of online behavior as harmful or criminal? Are there gendered nuances in this categorization of an online behaviour as a crime or harmful?
- RQ 2: What strategies and behaviors do young users employ to protect themselves online and where do privacy affordances fail them?
- RQ 3: What are their mechanisms for reporting or seeking support in a context like Pakistan which has a limited legal framework for the digital world?

Our research consisted of Repertory Grid (RGT) interviews with 18 participants (13M, 5F) to address RQ 1 and semi-structured qualitative interviews with 34 participants (17M, 17F) to explore RQs 2 and 3. Both studies had unique participants.

3.1 Repertory Grid Study

The study employs the Repertory Grid Technique (RGT) to elicit personal constructs and to observe the perception of participants regarding cyber threats. Developed by George Kelley as part of his Personal Construct Theory, it posits that people construe reality according to their personal constructs [10],

and they form these constructs by observing the contrasts between a set of examples [32]. The main components of RGT are *elements*, *constructs*, and *linking mechanisms* [50]. *Elements* represent objects of thought (people, places, ideas, or inanimate objects) that are compared methodically to discover the constructs of a person [22]. *Constructs* are the discriminations that people make between the elements. *Linking mechanisms* are the ways that show how participants interpret each element relative to each construct [50]. In our case, elements are cybercrime threats; constructs are characteristics that participants use to describe similarities and differences between cybercrime threats; and linking mechanisms are ratings of the threats made by the participants on each construct. Rating refers to the process of comparing or evaluating elements on each construct using a numerical or qualitative scale to capture individual perceptions and distinctions. A diagram explaining the methodology can be seen in Figure 1.

More precisely, we employed the Full RGT Method [44] where both the elements and the constructs were elicited from the participants. The participants were first asked about their personal experiences and the experiences of their close acquaintances with cybercrime to elicit the elements and finalize the cybercrime element list. After this element elicitation phase, the construct elicitation phase started where participants were presented with triads of elements organized in the triadic form [50] (see Table 3 in Appendix for triad order). Participants were instructed to compare and contrast any two most similar elements with the third one. To understand the underlying assumptions and reasoning behind the elicited constructs, the participants were further probed using "Why?" and "How?" questions (also called the Laddering Technique [26]) whenever needed. For instance, one of the presented triads during the study involved *hacking*, *unsolicited contact*, and *non-consensual use of information (NCUI)*. A participant mentioned that hacking and NCUI are similar, stating that they are more harmful compared to unsolicited contact. When inquired about the reason behind the choice, the participant explained that hacking and NCUI could potentially lead to the unauthorized access of personal pictures, which could then be used for blackmail. In contrast, unsolicited contact was seen as less directly harmful to the victim. The constructs were then recorded on a repertory grid (see Figure 4 in Appendix), and participants were asked to rate each element in relation to each (self-generated) construct using a five-point Likert scale (Linking Phase).

3.2 Qualitative Study Design

The study protocol consisted of 8 sections, which aimed to elicit information about the participants' device and internet usage and their experiences and beliefs regarding cybercrime and reporting mechanisms. To validate the protocol, pilot interviews were conducted, and the protocol was revised based on the findings from these interviews. To address the linguistic diversity of the participants, the protocol was translated into both English and Urdu. The average length of the interviews was approximately 0.7 hours, ranging from 0.31 hours to 1.2 hours. Sampling continued until data saturation was achieved, at which point no new information was obtained. Interviews were conducted online through Zoom. The interviews were conducted in a mixture of English and Urdu languages.

3.3 Participant Recruitment

The participants were recruited using a snowball sampling technique through personal contacts and online forms posted on university forums. The participants were pursuing degrees in various majors including STEM, Business, and Humanities. The interviews were conducted both in person and online on Zoom.

For the RGT study, the sample consisted of undergraduate students enrolled in 8 universities in Pakistan. A pilot study was conducted with a sample of 5 participants using the Repertory Grid Technique (RGT) method to establish a definitive methodology for the final interviews. We continued to recruit participants until saturation was reached, i.e. when no new threats or experiences of privacy violations emerged. A total of 18 participants (5 females, 13 males) with ages ranging between 18-24 were recruited from 8 universities in Pakistan. The demographics of the participants are presented in Table 1.

The qualitative study sample consisted of undergraduate students enrolled in 12 universities in Pakistan. A total of 34 participants (17 females, 17 males) with ages ranging between 18-24, were recruited from 12 universities in Pakistan. The demographics of the participants are presented in Table 4 in Appendix.

Gender	Male Female	13 5
Age (years old)	18 19 20 21 22 ≥ 23 Average, Median, Mode	1 3 2 6 2 4 20.94, 21, 21
Education Year-(Undergraduate)	Freshman Sophomore Junior Senior	2 4 3 9
University	Private Public	5 3

Table 1: RGT Demographics

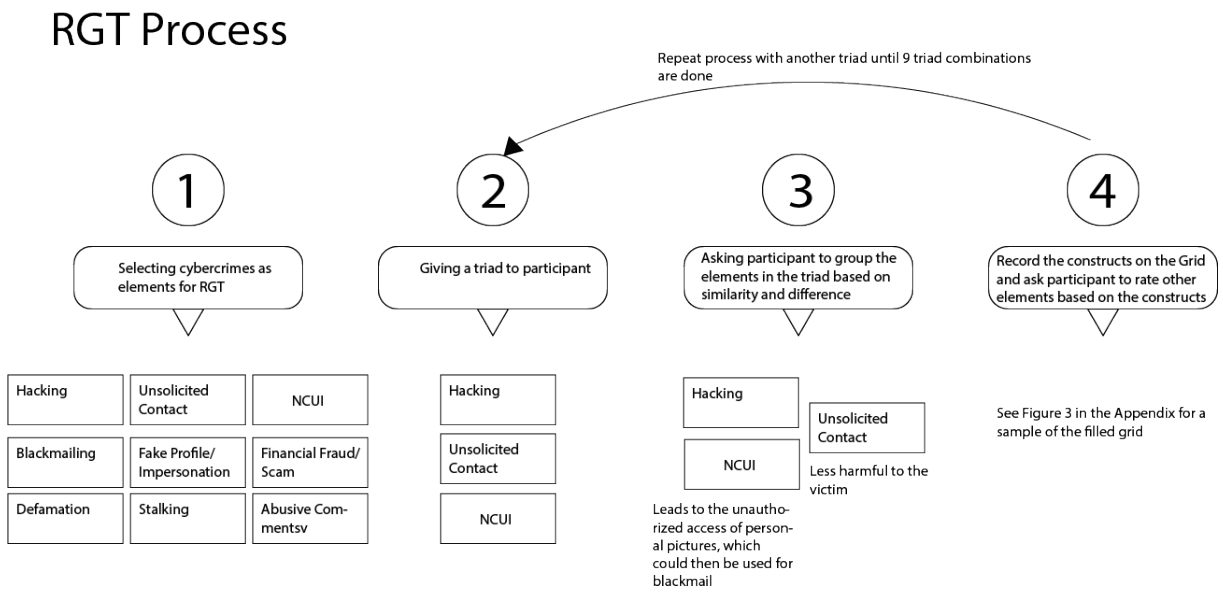


Figure 1: Methodology of the RGT process

3.4 Ethical Considerations

Both studies were approved by the IRB at the university where the study took place and informed verbal consent was obtained prior to conducting the research. The participants were informed that only audio recordings would be made and that the data would be used exclusively for research purposes. Additionally, it was communicated to the participants that the data would not be shared with any third parties and that any personally identifiable information would be removed during transcription to maintain anonymity.

At the time of the study, the minimum wage for unskilled workers and adolescent workers in Pakistan was PKR 120.2 (0.45 USD) per hour [5]. All participants were compensated for their time. The participants in RGT interviews were compensated PKR 500 (1.87 USD), while those participating in qualitative interviews were compensated PKR 1000 (3.74 USD). Participants were also informed that they could decline consent for recording without any impact on the compensation offered.

Given the sensitive nature of the subject matter and the cultural context of Pakistan, a female researcher was designated to conduct interviews with female participants, while male researchers were responsible for interviewing male participants.

3.5 Data Analysis

The recorded data was first transcribed. The data was then analysed using open-coding [31] which was conducted by a team of three researchers. To ensure consistency in the coding process, the first three interviews were collaboratively coded and an initial codebook was created. Subsequently, each researcher conducted individual coding, and recurring codes were consolidated during meetings. A total of 3,852 codes were generated from the transcripts, which were grouped into themes using Thematic Analysis, as described by Brown et al. [16]. The themes were further synthesized and organized using Affinity Mapping.

3.6 Positionality

The authors of this study are based in Pakistan and comprised of two female and two male researchers who were residing and working within the country during the time of the research. An additional 2 authors are based in Germany. Among the authors, two were considered to be young adults during the study, providing them with an advantage in their ability to relate to the experiences of the participants. This, in turn, facilitated an intuitive understanding of the social and religious context of the participants' responses pertaining to cybercrime. Additionally, the female researchers of the study have previously lived in both the US and Europe, and have spent much of their formative years in Pakistan which allows for a unique understanding of the Paksitani context.

4 Findings

Our findings highlight the cybercrime experiences, digital safety perceptions, safety behaviours, and available support systems of educated, tech-savvy users in Pakistan. We dis-

cuss the differences in these experiences and behaviours as compared to earlier reported experiences of low-literate, low-income users [39,46] in Pakistan and broadly the experiences and behaviours of young people in the Global North [13,37].

We find, as reported in prior work, phishing, fake profiles and impersonation, financial fraud and cyber-stalking to be frequently experienced harms in our context [9,33,38]. However, we also find specific gendered nuances and differences in what is considered a harm, how harms are experienced, the effectiveness of existing privacy affordances and the barriers to reporting that have not been reported in prior work.

We structure our findings below by first presenting a cybercrime spectrum which is based on the RGT study data and the qualitative data both (Section. 4.1). The rest of the findings are based only on the data from the qualitative study.

4.1 Cybercrime Spectrum: RGT Data and Qualitative Analysis

We used data from our RGT study along with qualitative data to calculate a dis-aggregated spectrum of cyber threats from least severe to most severe (Figures 2, 3). In the Repertory Grid Technique (RGT) interviews, multiple constructs were elicited to indicate the severity of each threat, such as Emotional harm (vs. Physical harm), No direct harm (vs. Direct harm), and Potentially harmful (vs. Less harmful). Figure 4 displays a sample of a grid from our RGT study. During the categorization process, elements were assessed based on their proximity to poles indicating severity or benignity. If an element was ranked towards the pole that indicated severity (i.e. More Severity, Potentially Harmful, Exploitation of the Victim), it was considered a severe threat. A similar process was applied for the poles (i.e. Mildly Threatening, Less Harmful, The Victim is not Exploited) that implied benignity. Similarly, during the qualitative interviews, participants were asked which online threats they considered severe and benign. We measured the frequency of each threat they rated severe or benign by grouping the responses. The total frequency of each threat was calculated by summing the frequencies of benign and severe threats obtained from both the qualitative and RGT interviews. Finally, the final severity rating for each threat was determined by subtracting the total benign frequency from the total severity frequency. Threats with higher scores were considered more severe, while those with lower scores were categorized as benign.

We see notable differences between male and female spectrums. Male participants considered Hacking, Blackmailing, NCUI, and Financial Fraud to be more severe while female participants considered Defamation, Hacking, NCUI, and Fake Profile to be more severe. Male and Female Participants both considered Stalking, Abusive Comments, and Unsolicited Contact to be less severe. Female participants also considered Financial Fraud to be less severe, which is in contrast to the male spectrum. In the subsections below we unpack and contextualize the spectrum based on the experiences and concerns expressed by the participants along with the prevalent socio-cultural norms based on our qualitative study.

4.2 Cybercrime Experiences and Concerns

Threat	Total (%)	Males (%)	Females (%)
Unsolicited Contact	20	2.5	17.5
NCUI	17.5	0	17.5
Hacking	15	7.5	7.5
Fake Profile & Impersonation	15	2.5	12.5
Financial Fraud/ ScamCalls	15	7.5	7.5
Blackmailing	7.5	0	7.5
Defamation	7.5	0	7.5
Stalking	2.5	0	2.5

Table 2: Frequency of threats reported by male and female participants

Participants in our qualitative study reported 40 personal cybercrime experiences, and 35 experiences of other people (family, friends, media coverage). Table 2 displays the frequency of each reported threat experienced by participants. The types of cybercrimes experienced by the participants, along with their definitions, can be found in Table 5 in Appendix. The definitions were supplemented from the Digital Rights Foundation [25], which is a Pakistani research-based advocacy NGO focusing on technologies to support human rights, democratic processes, and digital governance.

In the following subsections, we discuss the cybercrime experiences of five cyberthreats: unsolicited contact, cyberstalking, fake profile/ impersonation, financial fraud/ scam calls and non-consensual use of information. We highlight the concerns, the platform level affordances and participants own mitigation strategies for each crime. For each we also detail the reality of participants experiences. The following sections are based on data from our qualitative study.

4.2.1 Unsolicited contact

Unsolicited contact was the most frequently reported cybercrime across all female participants.

Concerns: Male and female participants considered unsolicited contact benign (relatively harmless crime). Female participants reported being contacted on various platforms without their consent, revealing that the privacy affordances provided by social media platforms were ineffective against unsolicited contact. Female participants would persistently receive message requests and calls despite blocking the accounts on social media and the contact numbers multiple times.

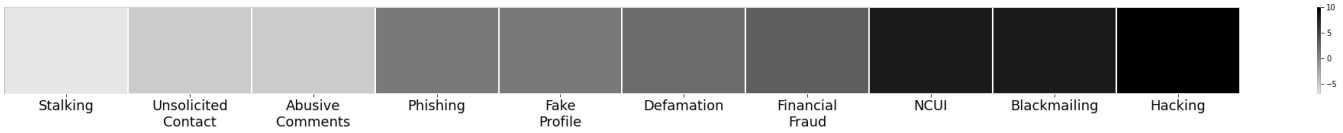


Figure 2: Male Spectrum about Cyberthreats

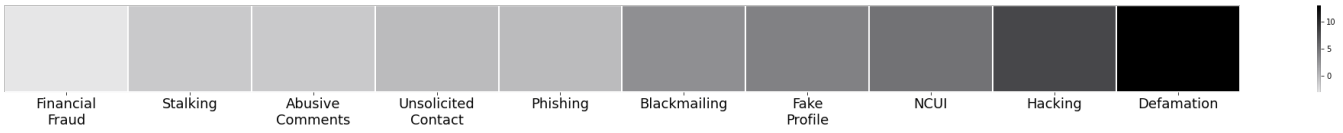


Figure 3: Female Spectrum about Cyberthreats

The perpetrators were always male, and their motive was to establish friendships with female users. One female participant highlighted: *"This [unsolicited contact] is such a common occurrence; I mean, people don't even call this [unsolicited contact] a cybercrime because it's that common. A random person texts you on Instagram and forces you to be friends; like, it's so common now that you don't even pay attention to it. You're just like this happens and stuff." - PIFT-F1*.

We found that users misuse the disappearing messages and one-time picture view feature of Snapchat to perpetrate unsolicited contact. Snapchat servers are designed to delete all Snaps (pictures) after all recipients have viewed them. Since the chats get deleted automatically, perpetrators send offensive content to users with the affirmation that the evidence of harassment will be erased permanently: *"Recently, in some harassment cases, we got to know that the harassers harass [others] on such platforms like Snapchat where all the chats are deleted... they don't want the chat to stay." - PF1*.

In another incident, one participant reported an instance of misuse of the Airdrop feature on an iPhone smartphone. The participant explained that their friend was traveling on a bus and forgot to turn off her Airdrop, which allowed a stranger to connect to her device and send unsolicited images.

Sambasivan et al.'s work in India, Pakistan, and Bangladesh highlights that 65% of the women in their sample reported friendship requests and unwanted phone calls from strangers as a common form of online abuse [46]. In contrast, unsolicited contact has not been reported as a frequent threat in the Global North [13, 15, 38]. Cultural norms around the segregation of genders and the importance placed on the modesty of women also impact the severity of some privacy violations as more traumatic in our context than others.

Affordances and Mitigation: Social media platforms, such as Instagram and Facebook, allow users to make their profiles private and block or report the abuser's profile. Participants take additional precautionary measures to ensure their privacy by not sharing their contact details with strangers and restricting the requesting profiles.

Reality: Despite platform level affordances, participants were unable to effectively navigate unwanted contact. Instagram's feature to allow users to create multiple accounts from a single profile was a contributing factor to the high prevalence of unsolicited contact on the platform. To address this issue, Instagram introduced a feature that allows users to block an account and any future accounts created using the same email address or contact number. This allows users to block contact with an individual's current account and any future accounts that the person may create in order to contact them again. However, participants found this feature to be ineffective in blocking unsolicited contact as users make new accounts with new email addresses: *"I even tried the option on Instagram to report all future accounts made by this person [the perpetrator], but maybe he made an account from a different email [that I kept receiving his messages]." - PU-F1*.

Participants also reported that the perpetrator often created a new account or phone number to contact them despite being blocked. This led to a sense of hopelessness among female participants, who had come to accept this type of cybercrime as a normal part of their online experience. On the other hand, unsolicited contact was not common among male participants, as evident from the Table 2, which may have contributed to their perception of it as benign. However, male participants were aware of its prevalence in society: *"The most frequent cyber-crimes you tend to hear about are messages to women, pictures of genitalia, or posts or texts mentioning lewd activities that they would like to do to the said women." - LM1*.

4.2.2 Cyberstalking

Our findings reveal that while both male and female participants found cyberstalking a benign threat, it was reported only by female participants.

Concerns: Cyberstalking was viewed as normal amongst our participants who did not consider it as *illegal* since social media platforms do not prevent users from accessing other

users' profiles. Participants understood public profiles as fair game for abusive comments and stalking without fear of legal repercussions. They believed it was perfectly legal to interact with profiles (in any way) as long as they were public. The participants being stalked considered it to be less severe as they were not *aware* of the fact they were being stalked: *"I would say stalking is immoral but not illegal since you are putting the information out there in public yourself" - LM4.*

We found male participants were relatively less concerned about cyberstalking than female participants. They believed they could ward off the stalkers through their physical strength. However, male participants were more concerned with online tracking and stalking by social media companies. They expressed concern about the platforms themselves spying on their digital activities. They explained that they were worried about being shown ads for something they only verbally discussed with their friends: *"Cyberstalking is also quite a threat, but only if the companies conduct it; stalking happens through [online] platforms." - LM2*.

Female participants believed they could be stalked through their laptop cameras and were concerned about hacking of their cameras and capturing of their photos in compromising positions. They strongly believed that cyberstalking often leads to crimes in the real world which include physical stalking and harassment: *"They [stalkers] get their [female victims] address and phone number, and they get into more detail about how they get their address and phone numbers. And then they follow her, making her extremely uncomfortable." - GF3*.

Studies in the region have not previously reported cyberstalking [39,46] as most studies have worked with low-literate populations. In contrast, our participants were young and tech-savvy, with much more engagement with online spaces and platforms.

Affordances and Mitigation: Mobile applications ask users before accessing multimedia, contacts, and camera of the user's device. Whatsapp allows its users to limit access to their profile picture and statuses to specific contacts. In addition, our study participants reported taping their front-view laptop cameras in fear of stalking by hackers. They avoided posting personal content, such as pictures and contact details, to the public.

Reality: We found that default privacy settings of social media platforms can significantly impact users' experience and the potential for unwanted or harmful interactions. Participants perceived Instagram to be more secure as compared to Facebook because when creating an account on Instagram, the default profile picture settings are set to private, while on Facebook, users are required to enable it manually. However, participants mentioned third-party external websites that can be used to enlarge and view anyone's profile pictures on Instagram. Originally, Instagram restricts users from enlarging the profile picture of any other user. These third-party links are easily accessible through online search engines. As far as we know, Instagram has failed to address this issue.

4.2.3 Fake profile & Impersonation

Concerns: Perpetrators created fake profiles using false information to exploit contacts and tarnish reputations. They frequently posed as women to gain access to women's accounts and then sent victims inappropriate content, such as explicit images and texts. In certain instances, they utilized real details of other individuals to impersonate them online. The motivation behind these impersonations often involved defaming the real individuals or establishing trust by pretending to be someone close to them.

A female participant reported: *"There was this guy who liked me. I didn't wanna get involved with him, so he got angry and sent me a friend request [through a fake account]. I thought it was my friend's and I accepted his request. He stole all my pictures with screenshots. Then he created another account and uploaded those pictures with captions that were not very pleasant." - FJMUF1*

Participants expressed deep concerns about the potential damage to their reputation caused by fake profiles. They emphasized the harm associated with a counterfeit profile impersonating them and sharing inappropriate or questionable content, leading others to mistakenly hold them responsible for it.

Affordances and Mitigation: Social media platforms allow users to create unique usernames, protecting against profile impersonation. Each profile is linked to a separate email address and mobile number, strengthening security measures. Moreover, users can report impersonating profiles and request the platform delete the profile.

To further protect themselves from fake profiles our study participants reported that they checked the requesting profile's activity (when a friend request is made) to verify its authenticity. To prevent the misuse of their display pictures, female participants in our study frequently blurred them.

Reality: Our research uncovered a contrasting reality. Perpetrators in our study possessed multiple SIM cards registered under their names, enabling them to create numerous profiles on social media platforms such as Instagram. In Pakistan, individuals can register up to five SIM cards using a single ID card. Similarly, Instagram permits users to create up to five profiles linked to a single email address.

Participants' lack of trust in official reporting mechanisms and cybercrime agencies compelled them to take matters into their own hands. Female participants, for instance, formed online groups to collectively report and flag fake accounts engaged in harassment. This approach proved effective as submitting a large number of reports within a short time frame increased the chances of the social media platform suspending the offending account. Additionally, they employed call-out posts to publicly shame perpetrators, recognizing that male wrongdoers were concerned about their social image and dam-

aging their reputation significantly. By publicly defaming the perpetrators on social media platforms, female participants effectively discouraged their harassing behavior and instilled fear among other men, deterring them from engaging in similar activities. : *"I just now put it in my friend's group and tell them to mass report it (the account), and their (perpetrator's) account gets disabled. You do name-shaming or call out a person (on social media). Tweet about them. Now people are scared to do such stuff because they know that if you tweeted about it and other people saw it, people are gonna suspend your account." - GCU-F2*.

4.2.4 Financial Fraud/ Scam Calls

Concerns: Scammers frequently targeted victims by assuming authoritative roles, such as bank employees or members of reputable community organizations. They utilized tactics like account-blocking threats or enticing cash rewards to obtain personal information. This information was then exploited for fraudulent transactions or to deceive victims into believing they had won lottery prizes, often requiring a small registration fee. However, our participants displayed a high level of awareness about prevalent scams in Pakistan and demonstrated the ability to recognize and avoid them easily.

Participants tended to blame the victims of scam calls and financial fraud, perceiving it as their own fault. Since the participants had never fallen victim to a scam, they only recounted scenarios they had observed. The victim in such scenarios was usually elderly, low-literate, and tech-illiterate: *"If a person is going to random sites and not verifying their authenticity, then it is their fault too. People should be careful themselves; you can't just blame the person committing the crime." - NUSTM1*

Affordances and Mitigation: Recent smartphone updates have introduced features that flag incoming calls from unknown numbers, aiding participants in our study in identifying potential scam calls. Through community-based reporting, if a phone number receives multiple spam reports, it can be labeled as a potential scam number, and new users will be notified accordingly. In addition, we observed that our participants employed various strategies to assess the authenticity of websites before engaging in transactions. For instance, they relied on indicators such as the site's popularity, product reviews, and visual aesthetics to establish a level of trust and convince themselves of the site's legitimacy.

Reality: No personal experiences of this cybercrime were reported, hence no vulnerabilities were identified.

4.2.5 Non-Consensual Use of Information

Concerns: NCUI (Non-consensual Use of Intimate Images) was exclusively reported by female participants in our study. Our findings established a clear connection between NCUI, fake profiles, and blackmail. Perpetrators gained unauthorized access to victims' personal data, which they then utilized to either blackmail the victims or create fraudulent profiles. *"There was some guy having my photos, he was basically blackmailing me into meeting him or else he'll get my photos and post them" - LCWU-F1*

The personal information was obtained either through the victims' social media accounts or, in one instance, through non-consensual dissemination by their friends. The perpetrators, predominantly males, would approach female participants under the pretext of initiating a relationship. *"She [the friend] gave my pictures to some person and then he texted me and is like I have your pictures, I know who you travel with in the school van; all you have to do is talk to me everyday. Otherwise, I will create a fake account using your pictures. GCU-F4*

Affordances and Mitigation: Snapchat's screenshot notification feature prevents the unauthorized use of personal photographs by alerting users when their snaps are captured. Our findings indicate that features providing more control over content visibility and lifespan enhance user experience and foster greater trust in the platform. Participants responded positively to Snapchat's timed snap feature, which allows users to set a viewing timer on their snaps, making them accessible for a specific period. This feature instills a sense of security, enabling users to share pictures without worrying about misuse, as recipients can only view them within a limited timeframe.

Reality: Despite some platform affordances, participants highlighted the ineffectiveness of Snapchat's screenshot notification feature when a user takes a screenshot by activating airplane mode on their mobile device, as it does not trigger a notification: *"If someone takes a screenshot [of the chat], you are notified, but even that has loopholes where people use it with airplane mode and stuff." - PIFT-M1*

Participants expressed concern about the limitations of the timed snap feature, as it was possible for users to capture and distribute the content using a different smartphone, discouraging them from sharing personal content through snaps. Similarly, on WhatsApp, blocking or deleting a contact does not delete the chat history between users, causing serious concerns among participants regarding potential chat leakage and the potential for their chats to be used against them: *"For instance, your conversation with someone comes to a close; even if you delete the stuff [chats], they will still have all the pictures downloaded in their phone. That is the only issue in Whatsapp." - LM2*

4.3 Barriers to Reporting Cybercrime

We found three major barriers when participants reported cybercrime to concerned authorities. These included the ineffectiveness of reporting platforms, lack of awareness regarding reporting mechanisms, and concerns about families.

4.3.1 Effectiveness of Reporting Platforms

Participants were skeptical of the effectiveness of social media platforms in resolving their reports of cybercrime. Additionally, they expressed a lack of trust in reporting such crimes to legal authorities. This contrasts with the experiences reported by US undergraduate students, as previously documented in the literature, where a greater level of comfort was reported in relation to reporting cybercrime to appropriate authorities [13].

The participants in our study revealed a lack of trust in the ability of legal agencies to effectively and efficiently resolve their reports of cybercrime. This sentiment was further reinforced by concerns about the potential for excessive information-gathering and the dissemination of sensitive personal information to third parties. Additionally, participants were concerned about legal agencies contacting their parents or gaining access to other personal information in the process of reporting cybercrimes. Overall, these concerns regarding privacy and the handling of sensitive personal information contributed to a reluctance to report cybercrime to legal agencies: *"If my email account is hacked, there is much more [information]. If that email account is connected to several other accounts, they [cybercrime agencies] will know which platforms and accounts I am using. They can access my data from those accounts." - LM2*.

Participants were also reluctant to report cybercrimes to social media platforms, citing the inefficiency of the platforms in taking timely action on complaints. They explained that the damage had already been inflicted on the victims by the time social media platforms took appropriate action toward the complaint. This is one of the reasons that people took matters into their hands: *"I have had my pictures used in contexts where I did not want them to be used. Someone started uploading my photos with crude captions wherein my response was to search online how to remove them through reporting systems on Instagram. And what I learnt from that was that by the time Instagram would sift through and decide it was worth removing, the damage would have been done. It would take less than twenty-four hours for me to become a laughingstock." - LM1*

In Pakistan, there is a general mistrust of government institutions, as individuals often encounter issues such as delayed or unresponsive responses, complex procedures, and uncooperative staff when interacting with these departments. This mistrust extends to Pakistan's Federal Investigation Agency (FIA) specifically when it comes to reporting cybercrime. Participants noted that the FIA does not provide statistics on the number of crimes resolved, making it difficult for them to assess the agency's efficiency. There is also little transparency about the process of lodging a complaint or what happens once a complaint has been made.

However, a few male participants told us that the agencies working against cybercrime in Pakistan were doing a satisfactory job. They explained that once a crime is reported to FIA, they resolve it effectively and promptly: *"Yes, FIA is doing a good job because once you complain to them, they take two days max to reply to you." - LM3*.

It is important to note that the participant only mentions the time taken to receive a reply from FIA. Resolving a complaint takes an even longer time. In contrast, female participants mentioned that FIA is not helpful in the majority of the cases. Most female participants did not report cybercrimes they faced to any legal agency, citing a lack of knowledge about whom to contact and how to report such crimes. One participant provided an example of a case of blackmail on Facebook involving another girl, in which her pictures were leaked, and she was being blackmailed for a large sum of money. She explains that the FIA was not helpful in resolving this crime.

Similarly, social media platforms do not take contextualized action against the reported cybercrimes. One participant reported that a fake account was made using her name on Facebook. The perpetrator blocked the participant from the fake profile. Despite consistent requests to Facebook and cybercrime agencies, the participant could not get the fake account deleted. Expressing concern over the non-consensual use of her pictures on a fake account, the participant mentioned: *"In what I experienced, the [fake] account was not disabled and it had about 500 people in the friend list. Using my pictures, I did not know whom they were talking to or what they were talking about. My concern is that somebody is using my identity, that is why I am very concerned about my pictures that they do not get leaked anywhere." - GCU-F4*. In contrast to formal pathways, participants often preferred utilizing their personal contacts in cybercrime agencies so their reports could be heard and appropriate action could be taken against the perpetrators. Similarly, perpetrators from influential families often use their political connections to intimidate the victim to not report cyber-crimes. Participants also expressed concern that if they report the cybercrime and the perpetrator finds out, they could make their life more difficult if they had such connections.

These concerns and the general lack of transparency in the procedures and mechanisms for how platforms and local agencies handle complaints leads to an absence and vacuum of support mechanisms for users in Pakistan.

4.3.2 Lack of Awareness Regarding Reporting Mechanisms

Along with distrust in cybercrime agencies, another important barrier when reporting cybercrime for our participants was the lack of awareness and education regarding reporting mechanisms. The participants were unaware of agencies working to curb cybercrime. Participants mentioned that they would only contact agencies if they could not solve the problems themselves or with the help of their friends. Only a few partic-

ipants were aware of the Federal Investigation Agency (FIA) as a possibility for reporting crimes. In general, female participants were not aware of where to report. This lack of awareness was also cited as a reason why women in Pakistan do not report cybercrimes. When asked how they would report a cybercrime, participants believed the reporting procedure to be complicated and beyond their expertise: *"I don't think I'll take any steps to report cybercrime because it's quite complicated, and I do not know where to report it and what the procedure is. I have no idea." - FJMU-F1*.

Our participants were aware of the reporting features of the social media platforms they used, which contrasts with what Sambasivan reported [46]. However, they did not find the response from the platforms to be appropriate enough to deal with the cybercrime (more in Section 5.4).

We also found the sources of awareness regarding cybercrimes and privacy among young adults in Pakistan, which differed from the sources previously reported amongst low-literate populations in Pakistan as reported by Naveed et al. [39] but are more similar to the sources reported amongst US populations [41]. These sources were:

1. Friends: Participants, both male and female, reported that they mainly learned about cybercrimes and privacy features of applications from their friends.

2. Social Media Groups: Participants, primarily female, reported that they learned about cybercrimes prevalent in Pakistan through social media posts. They explained that they had joined groups on Instagram or Facebook where posts about such topics were made.

4.3.3 Concerns about Families while Reporting

Participants expressed concerns regarding family reactions when it came to reporting. They preferred not to inform their family about experienced cybercrime. They explained that if the family got aware of the cybercrime situation, they would start worrying, and it would cause them mental stress. One participant mentioned that if someone has not done anything wrong, they should tell their family about the cybercrime. However, what constitutes as wrong varies from family to family. Since the burden of maintaining family's honor often falls on women in Pakistan [39], even talking to men online could be considered a wrongful act by women. Noting this, female participants expressed concerns about being victim blamed if they informed their family members about the cybercrime violation. They believed their parents would not be supportive of their actions and would point out faults in their actions. Victim blaming is also common in Pakistan, especially regarding women. Participants expressed that they do not have enough space to talk about these issues safely: *"Because the females are being affected by cybercrime so much that we cannot even talk about it. And whoever does, gets victim blamed that it's your own fault. That's the main problem that we don't get enough space to talk about it or be heard." - GF3*.

Additionally, family members discourage female participants from reporting cybercrime to concerned authorities and ask them to instead block the perpetrator and ignore it. This is typically done to preserve family honour and reputation within their social circles. Due to this, participants preferred resolving cybercrime situations personally to contain its spread to family members or the public. Such familial concerns have not previously reported as barriers in the Global North [13, 15, 19, 37].

5 Discussion

Our work examines experienced and perceived online harms and cybercrimes within Pakistan's educated and technologically proficient young adult population. The security gaps discussed in section 4.2 have serious implications in a complex context like Pakistan, particularly for young people who navigate religious values, family honour, peer pressure, a lack of legal support and gendered expectations in online spaces.

We highlight key insights from our findings below:

- We find distinct gendered differences in the experiences of and types of harm from cyber-crimes, with female users predominantly harmed through violations negatively impacting their reputations or those of their families like defamation, fake profiles or NCUI. In contrast, male users are often more concerned with financial frauds often because in Pakistan they are responsible for finances and actively conduct financial transactions.
- Significant emphasis is placed on social standing within the community in this context and so users are reluctant to report cyber-crimes or seek help from authorities. There are also few legal frameworks tackling cybercrimes, leaving young people with little support.
- Users are very aware of platform level vulnerabilities, but often not of platform affordances. This coupled with an inadequate response from reporting to platforms means they often rely on non-technical (social) mechanisms to protect themselves. For example, female users engage in collective, mass reporting of accounts used for harassing other female users (within a short period of time) to shut down the account.

It is important to highlight here that most often the focus in South Asia is on privacy *literacy* as most prior work in the region has focused on low-literate or low-income populations [15, 19, 39, 45–47]. In contrast, we find even with tech-savvy, literate and early adopters of technologies, platform privacy features fail to provide contextualized privacy affordances.

Our findings in particular highlight the gendered differences in how cybercrimes are experienced and perceived. While this is also reported in studies in the US and UK [34,55], we find in the Global South context (India, Bangladesh and

Pakistan), female users report a higher incidence of cybercrimes like unsolicited contact, non-consensual use of information (NCUI), fake profiles and impersonation [46]. These similarities suggest a shared cybercrime landscape among these countries. In contrast, identity theft was a greater concern for individuals in the USA, while hidden costs in services, frauds, and scams were more worrisome for Germans [29]. Shoulder surfing, a threat not found in our target demographic, raised concerns among individuals in Germany and Saudi Arabia [45]. Our work in addition to prior work in Pakistan, India and Bangladesh [18, 20, 39, 46, 47] suggests some shared experiences, concerns, harms and mitigation strategies across all these countries, highlighting the need for culture, context specific privacy design.

5.1 Design Implications

Based on our data we identify several design opportunities for addressing the concerns of our population. Despite their technical proficiency, our participants demonstrated a lack of knowledge about the privacy features provided by social media applications. The results of our study indicate that multiple participants were not aware of the privacy affordances provided by social media platforms that made them vulnerable to cybercrimes. One possible way to address this issue is to use geo-location tagging to identify users in contexts where they might be vulnerable to specific privacy violations. Using this context based on geo-location, platforms could customise on-boarding procedures. Platforms should also consider switching from an opt-out mechanism for privacy settings to an opt-in default approach, whereby privacy preservation is the default setting. Below we propose mechanisms to counter the cybercrimes discussed in Section 4.2:

1. **Unsolicited contact**: One possible mechanism to tackle this is for social media platforms to consider implementing default settings that disable contact by strangers or provide users with the option to make these choices during an context-specific on-boarding process.

2. **Cyberstalking**: We propose that social media platforms notify users when their profiles are repeatedly visited by another user within a short timeframe. We also recommend that profiles should be locked by default or locked during privacy on-boarding.

3. **Fake profile**: Our participants identified a significant concern when reporting incidents to social media platforms, specifically their lack of visibility into the progress and outcome of their reports. This lack of transparency, particularly in cases where the reported incident was time-sensitive, such as defamation, led to participants attempting to resolve the issue on their own. To address this issue, we propose that social media platforms should implement a feature that provides users with a timeline of the progress of their reports. This would enable users to have greater visibility into the actions taken in response to their reports, and to make more informed decisions about their next steps. It is also vital that platforms create context aware, culturally appropriate guidelines to address reports. Geo-locations of the users reporting can be used to send the reports to specific channels to handle them with the relevant cultural context.

4. **Financial Fraud**: To enhance awareness and protect users from financial fraud, we recommend that social platforms implement a nudging strategy by regularly providing information about common scams in the user's country. By utilizing geolocation data, platforms can tailor the information to be specific and relevant to each user's location, helping them stay informed and vigilant against prevalent scam patterns in their area.

6 Conclusion

Our study employs the repertory grid technique and qualitative interviews to unpack users' mental models of cybercrimes, their experiences with cybercrime, and their privacy-preserving behaviors. We highlight the importance of understanding and incorporating specific cultural and religious values into the design to allow diverse users to freely use online spaces. We also underscore the challenges of designing in such nuanced and complex contexts where religious, familial, and cultural values often clash with user desires and online behaviours. Despite these challenges, it is important for designers and platforms to consider potential mechanisms to address the safety of young online users in contexts like Pakistan, where there is little legal support from local agencies.

Acknowledgments

This research was funded by the German Academic Exchange Service's (DAAD) "Deutsch-Pakistanische Forschungskooperationen" program under project ID: 57609539.

References

[1] Covid-19 exploited by malicious cyber actors-2020. https://www.cisa.gov/uscert/ncas/alerts/aa20-099a (02/05/2023).

[2] Digital development dashboard. https://www.itu.int/en/ITU-D/Statistics/Dashboards/Pages/Digital-Development.aspx (02/05/2023).

[3] Digital rights foundation, annual report 2021. https://digitalrightsfoundation.pk/wp-content/uploads/2022/05/helpline-annual-report-2021-1.pdf (02/15/2023).

[4] Interpol report shows alarming rate of cyberattacks during covid-19.

[5] Minimum wage notifications. https://efp.org.pk/minimum-notifications/.

[6] Ayaz Hussain Abbasi. Pandemic effect: Cybercrime on the rise. *T-Magazine*, 2022.

[7] Norah Abokhodair and Sarah Vieweg. Privacy & social media in the context of the Arab Gulf. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems*, pages 672–683, 2016.

[8] Tanisha Afnan, Yixin Zou, Maryam Mustafa, Mustafa Naseem, and Florian Schaub. Aunties, strangers, and the FBI: Online privacy concerns and experiences of Muslim-American women. In *Eighteenth Symposium on Usable Privacy and Security (SOUPS 2022)*, pages 387–406, 2022.

[9] Ahmed Aleroud and Lina Zhou. Phishing environments, techniques, and countermeasures: A survey. *Computers & Security*, 68:160–196, 2017.

[10] Antonia Bauman. The use of the repertory grid technique in online trust research. *Qualitative Market Research*, 18(3):362–382, 2015.

[11] Steven Bellman, Eric J Johnson, Stephen J Kobrin, and Gerald L Lohse. International differences in information privacy concerns: A global survey of consumers. *The Information Society*, 20(5):313–324, 2004.

[12] Noam Ben-Asher, Niklas Kirschnick, Hanul Sieger, Joachim Meyer, Asaf Ben-Oved, and Sebastian Möller. On the need for different security methods on mobile phones. In *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services*, pages 465–473, 2011.

[13] Morvareed Bidgoli, Bart P Knijnenburg, and Jens Grossklags. When cybercrimes strike undergraduates. In *2016 APWG Symposium on Electronic Crime Research (eCrime)*, pages 1–10, 2016.

[14] Timm Böttger, Ghida Ibrahim, and Ben Vallis. How the internet reacted to COVID-19: A perspective from Facebook's edge network. In *Proceedings of the ACM Internet Measurement Conference (IMC '20)*, pages 34–41, 2020.

[15] Casey Breen, Cormac Herley, and Elissa M Redmiles. A large-scale measurement of cybercrime against individuals. In *CHI Conference on Human Factors in Computing Systems*, pages 1–41, 2022.

[16] Nela Brown and Tony Stockman. Examining the use of thematic analysis as a tool for informing design of new family communication technologies. In *27th International BCS Human Computer Interaction Conference (HCI 2013)*, pages 1–6, 2013.

[17] David Buil-Gil, Fernando Miró-Llinares, Asier Moneva, Steven Kemp, and Nacho Díaz-Castaño. Cybercrime and shifts in opportunities during COVID-19: a preliminary analysis in the UK. *European Societies*, 23(sup1):S47–S59, 2021.

[18] Jay Chen, Michael Paik, and Kelly McCabe. Exploring internet security perceptions and practices in urban Ghana. In *10th Symposium On Usable Privacy and Security (SOUPS 2014)*, pages 129–142, 2014.

[19] Cassandra Cross. No laughing matter: Blaming the victim of online fraud. *International Review of Victimology*, 21(2):187–204, 2015.

[20] Jayati Dev, Sanchari Das, and L Jean Camp. Understanding privacy concerns of whatsapp users in india: poster. In *Proceedings of the 5th Annual Symposium and Bootcamp on Hot Topics in the Science of Security*, pages 1–1, 2018.

[21] Maeve Duggan. Online harassment 2017. 2017.

[22] Mark Easterby-Smith. The design, analysis and interpretation of repertory grids. *International Journal of Man-Machine Studies*, 13(1):3–24, 1980.

[23] Anja Feldmann, Oliver Gasser, Franziska Lichtblau, Enric Pujol, Ingmar Poese, Christoph Dietzel, Daniel Wagner, Matthias Wichtlhuber, Juan Tapiador, Narseo Vallina-Rodriguez, et al. A year in lockdown: How the waves of COVID-19 impact internet traffic. *Communications of the ACM*, 64(7):101–108, 2021.

[24] Jerry Finn. A survey of online harassment at a university campus. *Journal of Interpersonal Violence*, 19(4):468–483, 2004.

[25] Digital Rights Foundation. Helpline annual report 2021, 2022.

[26] Bannister Fransella, Bell. *A Manual for Repertory Grid Technique*. John Wiley, New York, 2004.

[27] Imran Gabol and Taser Subhani. Qandeel baloch murdered by brother in Multan: police. *DAWN.COM*, Jul 2016.

[28] Marian Harbach, Alexander De Luca, Nathan Malkin, and Serge Egelman. Keep on lockin' in the free world: A multi-national comparison of smartphone locking. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 4823–4827, 2016.

[29] Marian Harbach, Sascha Fahl, and Matthew Smith. Who's afraid of which bad wolf? A survey of it security risk awareness. In *2014 IEEE 27th Computer Security Foundations Symposium*, pages 97–110, 2014.

[30] Julio Hernandez-Castro and Eerke Boiten. Cybercrime prevalence and impact in the UK. *Computer Fraud & Security*, 2014(2):5–8, 2014.

[31] Judith A Holton. The coding process and its challenges. In Kathy Charmaz and Antony Bryant, editors, *The Sage handbook of grounded theory*, volume 3, pages 265–289. Sage, Los Angeles, 2007.

[32] Devi Jankowicz. *The Easy Guide to Repertory Grids*. John Wiley, New York, NY, USA, 2003.

[33] Katharina Krombholz, Dieter Merkl, and Edgar Weippl. Fake identities in social media: A case study on the sustainability of the Facebook business model. *Journal of Service Science Research*, 4:175–212, 2012.

[34] Carsten Maple, Emma Short, and Antony Brown. Cyberstalking in the United Kingdom: An analysis of the ECHO pilot survey. Technical report, University of Bedfordshire, 2011.

[35] Tara Matthews, Kathleen O'Leary, Anna Turner, Manya Sleeper, Jill Palzkill Woelfer, Martin Shelton, Cori Manthorne, Elizabeth F Churchill, and Sunny Consolvo. Stories from survivors: Privacy & security practices when coping with intimate partner abuse. In *Proceedings of the 2017 CHI conference on human factors in computing systems*, pages 2189–2201, 2017.

[36] Nora McDonald and Andrea Forte. The politics of privacy theories: Moving from norms to vulnerabilities. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2020.

[37] Matti Näsi, Petri Danielsson, and Markus Kaakinen. Cybercrime victimisation and polyvictimisation in Finland—Prevalence and risk factors. *European Journal on Criminal Policy and Research*, pages 1–19, 2021.

[38] Matti Näsi, Atte Oksanen, Teo Keipi, and Pekka Räsänen. Cybercrime victimization among young people: A multi-nation study. *Journal of Scandinavian Studies in Criminology and Crime Prevention*, 16(2):203–210, 2015.

[39] Sheza Naveed, Hamza Naveed, Mobin Javed, and Maryam Mustafa. "Ask this from the person who has private stuff": Privacy perceptions, behaviours and beliefs beyond weird. In *CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2022.

[40] Kaja Prislan, Igor Bernik, Gorazd Meško, Rok Hacin, Blaž Markelj, and Simon LR Vrhovec. Cybercrime victimization and seeking help: A survey of students in Slovenia. In *Proceedings of the Third Central European Cybersecurity Conference*, pages 1–2, 2019.

[41] Elissa M Redmiles, Sean Kross, and Michelle L Mazurek. How I learned to be secure: A census-representative survey of security advice sources and behavior. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 666–677, 2016.

[42] Carin MM Reep-van den Bergh and Marianne Junger. Victims of cybercrime in Europe: A review of victim surveys. *Crime Science*, 7(1):art. 5, 2018.

[43] Jake Reichel, Fleming Peck, Mikako Inaba, Bisrat Moges, Brahmnoor Singh Chawla, and Marshini Chetty. 'I have too much respect for my elders' understanding South African mobile users': Perceptions of privacy and current behaviors on Facebook and WhatsApp. In *Proceedings of the 29th USENIX Conference on Security Symposium*, pages 1949–1966, 2020.

[44] Ronit Rozenszajn, Galia Zer Kavod, and Yossy Machluf. What do they really think? the repertory grid technique as an educational research tool for revealing tacit cognitive structures. *International Journal of Science Education*, 43(6):906–927, 2021.

[45] Mennatallah Saleh, Mohamed Khamis, and Christian Sturm. What about my privacy, Habibi? Understanding privacy concerns and perceptions of users from different socioeconomic groups in the Arab World. In *IFIP Conference on Human-Computer Interaction (INTERACT 2019)*, pages 67–87. Springer, 2019.

[46] Nithya Sambasivan, Amna Batool, Nova Ahmed, Tara Matthews, Kurt Thomas, Laura Sanely Gaytán-Lugo, David Nemer, Elie Bursztein, Elizabeth Churchill, and Sunny Consolvo. "They don't leave us alone anywhere we go": Gender and digital abuse in South Asia. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, page 2, 2019.

[47] Nithya Sambasivan, Garen Checkley, Amna Batool, Nova Ahmed, David Nemer, Laura Sanely Gaytán-Lugo, Tara Matthews, Sunny Consolvo, and Elizabeth Churchill. "Privacy is not for me, it's for those rich women": Performative privacy practices on mobile phones by women in South Asia. In *Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018)*, pages 127–142, 2018.

[48] Express News Service. Girl commits suicide after morphed pics appear on Facebook... *The New Indian Express*, Jun 2016.

[49] Calvin Shivers. Covid-19 fraud: Law enforcement's response to those exploiting the pandemic. *Federal Bureau of Investigation*.

[50] Felix B. Tan and M. Gordon Hunter. The repertory grid technique: A method for the study of cognition in information systems. *MIS Quarterly*, 26:39–57, 2002.

[51] Kurt Thomas, Devdatta Akhawe, Michael Bailey, Dan Boneh, Elie Bursztein, Sunny Consolvo, Nicola Dell, Zakir Durumeric, Patrick Gage Kelley, Deepak Kumar, et al. Sok: Hate, harassment, and the changing landscape of online abuse. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 247–267. IEEE, 2021.

[52] Kurt Thomas, Patrick Gage Kelley, Sunny Consolvo, Patrawat Samermit, and Elie Bursztein. "It's common and a part of being a content creator": Understanding how creators experience and cope with hate and harassment online. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, page art. 121, 2022.

[53] Steve van de Weijer, Rutger Leukfeldt, and Sophie Van der Zee. Reporting cybercrime victimization: determinants, motives, and previous experiences. *Policing: An International Journal*, 43(1):17–34, 2020.

[54] Jessica Vitak, Kalyani Chadha, Linda Steiner, and Zahra Ashktorab. Identifying women's experiences with and strategies for mitigating negative effects of online harassment. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 1231–1245, 2017.

[55] Emily A Vogels. The state of online harassment. *Pew Research Center*, 13, 2021.

[56] S Elizabeth Wick, Craig Nagoshi, Randy Basham, Catheleen Jordan, Youn Kyoung Kim, Anh Phuong Nguyen, and Peter Lehmann. Patterns of cyber harassment and perpetration among college students in the united states: A test of routine activities theory. *International Journal of Cyber Criminology*, 11(1):24–38, 2017.

A RGT Protocol

Hello, thank you so much for agreeing to this interview! I am from <institution name>, and I'm working with my fellow researchers to understand the cybercrime experiences, perceptions, and understandings of young adults in Pakistan. In this interview, we hope to learn more about your digital activity and your experiences with cybercrime, if any. To accomplish this task, we will use an interesting interview technique called Repertory Grid Technique. I will explain the specifics of the methodology as we proceed.

Here are a couple of pointers before we start:

- We will compensate you PKR 500 for your time. Kindly share your account details at the end of the interview.
- I would ideally want to record this interview so I can later analyze your responses. Do you give me consent for the audio recording of this interview?
- We will keep your data anonymous and secure. It would not be shared with anyone apart from our research team. If your quotes are used in the final report, we will label the quote with a dummy label that cannot be traced back to you.
- The interview will take approximately 1 hour of your time. If you want to stop the interview at any point during this session, please let me know. You will still be fully compensated for your time.

If you have questions, then do let me know. I am going to start recording now. I would like you to reconfirm that you have given me consent to record this interview.

A.0.1 Focus: Demographics

- What is your age?
- What is your gender?
- What is your current education?
- What is your current occupation, if any?

A.0.2 Focus: Electronic usage

- How many electronic devices do you own?
- How many of the electronic devices you mentioned are shared among your friends or family?
- What are your most frequently used applications?

A.0.3 Focus: Opening questions

- Have you ever been the victim of a cybercrime? If so, can you tell me about your experience?
- Have you ever had to report a cybercrime to law enforcement? How did that experience go?

A.0.4 Focus: Methodology familiarization

Thank you for sharing your experiences. I will now introduce you to the Repertory Grid Technique. Let me walk you

through an example, so it is easier for you to understand the process.

- Tell me the names of any three Professors you have had the opportunity to work or study with.
- Can you tell me a way in which any two of these Professors are different from the third? Why is that?

I will write down your comparison on the grid now. On the left is the 'similarity pole,' meaning the property you found similar in two Professors. On the right is the 'contrast pole,' the property you found contrasting in the third Professor. [Details: A sample of the grid is shown in Figure 4]

- On a scale from 1 to 5, rate each Professor based on his/her closeness to the similarity or contrast pole. 1 means the Professor strongly lies in the similarity pole category; 5 means the Professor strongly lies in the contrast pole category. The middle value of 3 means the Professor cannot be classified in either of the categories or can be equally classified in both of them.
- Please justify your ratings for each Professor.

A.0.5 Focus: Element familiarization

Let's move on to the main part of the interview. Here is the list of 9 cybercrimes. In addition to these, I am adding cybercrimes you have personally experienced but are not on this list.

- Give me a definition of each of these cybercrimes. If I feel you are missing any crucial point, I will correct you.

A.0.6 Focus: Main Repertory Grid study

Great! We are all set. I will be presenting you with a random set of three cybercrime names one by one. You are required to compare and contrast any two of them with the third one. The process will be the same as the example we went through about the Professors. [Details: The triads were presented in the order shown in Table 3]

Triad No	Element 1	Element 2	Element 3
1	Hacking	NCUI	Unsolicited Contact/ Inappropriate Contact
2	NCUI	Unsolicited Contact/ Inappropriate Contact	Blackmailing
3	Unsolicited Contact/ Inappropriate Contact	Blackmailing	Fake Profile/ Impersonation
4	Blackmailing	Fake Profile/ Impersonation	Scam/ Financial Fraud
5	Fake Profile/ Impersonation	Scam/ Financial Fraud	Defamation
6	Scam/ Financial Fraud	Defamation	Stalking
7	Defamation	Stalking	Abusive Comments
8	Stalking	Abusive Comments	Hacking
9	Abusive Comments	Hacking	Scam/ Financial Fraud

Table 3: Triads

B Qualitative Study Protocol

B.0.1 Focus: Demographics

- What is your age?
- What is your gender?
- What is your current education?
- What is your current occupation, if any?
- What is your marital status?

B.0.2 Focus: Electronic usage

- How many electronic devices do you own? Name them.
- How long have you owned a device and have been using internet services?
- How many of the electronic devices you mentioned are shared among your friends or family? What purpose do they use your device for?
- What do you usually use your devices for? How many applications do you use? What are your most frequently used applications?

B.0.3 Focus: Internet consumption

- What do you think is the greatest privacy risks on the online platforms that you use?
- Do you share different information on different online platforms [including social media sites and e-commerce]? Why is it so?
- What kind of information (that you share online) is riskier and needs to be protected more securely?

B.0.4 Focus: Privacy-preserving mechanisms

- On a scale of 1 to 10 (1 being the lowest and 10 being the highest), how concerned are you about privacy violations?

- Which threats do you fear the most? Why is that? What measures do you take to protect yourself against them?
- Do you use any security measures to protect the data on your phone, device, and applications? If yes, what measures do you take?

B.0.5 Focus: Understanding of and experiences with cybercrime

- How would you define cyberspace?
- What do you think is a privacy violation in the digital space? What types of these violations are included in your interpretation of cybercrime?
- What demographics/ groups are more vulnerable to cyber threats you have mentioned? How can these demographics better protect themselves from these threats?
- Have you ever experienced a privacy infringement? If yes, what exactly happened? If not, do you know of anyone else who has experienced one? Explain.
- Do you think all cybercrimes are strictly punishable? Are there any threats that you think are unethical but not a crime?

B.0.6 Focus: Awareness of cybercrime

- Do you think cybercrime is increasing? If yes, what might be the reasons?
- What do you think can be done to control (handle) privacy violations/ cybercrime? [follow up on the answer; ask how and why?]
- Where do you educate yourself about (a) cybercrime, (b) Online ethics, (c) Privacy violations?
- What barriers have you faced in educating yourself regarding (a) cybercrime, (b) Online ethics, and (c) Privacy violations?

B.0.7 Focus: Cybercrime reporting

- If a cybercrime incident were to happen to you (being hacked/unauthorized data access), what would be your first step?
- Would you be comfortable reporting a cyber threat? If so, how and where would you report? and what would your expectation be (in terms of resolution)?
- Do you think that government organizations are playing their role actively in apprehending the perpetrators of cybercrime?

B.0.8 Questions related to individual threats [From RGT data]

- Which threats are more frequent in cyberspace, and what factors make them more frequent? Discuss the features of the threat which make them easy to perform.
- What factors make the identification (and reporting) of a threat difficult for the victim?
- Do you think cyber threats lead to threats in the physical space? How so?
- Why do you think someone would carry out a cybersecurity breach?
- Do you think there are threats where the victim is at fault for falling victim? Could they have been avoided by taking better precautionary measures?

C Qualitative Demographics

Gender	Male Female	17 17
Age (years old)	18 19 20 21 22 ≥ 23 Average, Median, Mode	3 2 5 14 7 3 20.85, 21, 21
Education Year (Undergraduate)	Freshman Sophomore Junior Senior	4 6 3 17
University	Private Public	5 7
Owned Devices	Smartphone Laptop Tablets Tv, PC, Console	34 32 5 5
Internet Usage (years)	1-5 6-10 11-15 16-20 Average, Median, Mode	9 7 7 3 9.48, 10, 5

Table 4: Qualitative Demographics

D Repertory Grid

1	Hacking	Unsolicited Contact/ Inappropriate Content	Non Consensual Use of Information	Blackmailing	Fake Profile/ Impersonation	Scam/ Financial Fraud	Defamation	Stalking	Abusive Comments	5	Participants
Mildly Threatning	5	2	3	5	4	2	3	1	2	More severness	IBA-M2
Potentially harmful (reputation and financial)	1	4	1	1	1	1	1	5	4	Less harmful (emotional distress only)	Fast-M2
Exploitation of the victim	1	4	1	1	1	1	1	1	1	The victim is not exploited	Lums-F2
Happens forcefully	1	1	1	1	1	1	1	4	4	Does not involve the use of force	Lums-F2
Victim's Personal information is not exposed	5	1	4	3	4	4	2	1	1	Victim's Personal information is exposed	IBA-M2

Figure 4: Repertory Grid - A sample of 5 constructs elicited from 4 different participants. The left column represents the similarity pole, and the second-last right column represents the contrast pole. The middle columns represent the various cybercrimes that the participants rated on a scale of 1 to 5

E Cybercrime Definitions

Cybercrime	Description
Hacking	Gaining unauthorized access to someone's electronic system, data, account, and devices, which can result in loss of data, loss of identity, and blackmailing.
Unsolicited contact	Unsolicited contact involves unwanted and repeated calls and messages by the accused/abuser, which may include spam, repeated requests for contact, personalized threats, blackmail, or any unwanted contact that makes the receiver feel uncomfortable.
Non-Consensual Use of Information (NCUI)	NCUI occurs when an abuser uses the victim's information without their consent and usually, without their knowledge
Blackmailing	Blackmailing involves using personal information or psychological manipulation to make threats and demands from the victim.
Fake Profile	Fake profile on a social media platform is an account pretending to be someone that does not exist.
Impersonation	When someone is using someone else's identity online and is acting as them online. It manifests in profiles purporting to belong to someone on social media websites and contacting people through texts or calls pretending to be someone else
Scam Calls/ Messages	Fraudulent calls that pretend to be an individual or from an authority to make a quick profit. Mostly such scam calls lead to potential financial fraud being committed.
Defamation	Defamation involves any intentional, false communication purporting to be a fact that harms or causes injury to the reputation of a person
Stalking	Stalking is keeping track of someone's online activity, without their knowledge, in a way that it makes the subject of the stalking uncomfortable.
Abusive Comments	Abusive comments involve the usage of harsh, hurtful, explicit, or insulting language to attack another person.

Table 5: Cybercrime definitions - The definitions were supplemented from the Digital Rights Foundation's 2021 Annual Report [25].

F Codebook of Qualitative Study

Top-level category	Description	Codes
Cybercrime perceptions	Subjective experiences, beliefs, and personal definitions of cybercrime.	1. Privacy violations 2. Associated risks
Privacy risks	Cybercrime threats through technology and their consequences.	1. Fears 2. Concerns 3. Online activities 4. Vulnerable demographics 5. Device sharing 6. Avoidance
Privacy control	Management of privacy on online tools.	1. Privacy affordances 2. Privacy-preserving practices 3. Private profiles 4. Two-factor authentication 5. Encryption 6. Screenshot notifications 7. Limiting account access
Reporting	Underlying challenges to reporting cybercrime incidents.	1. Hurdles 2. Family support/ resistance 3. Reporting venues 4. Nepotism 5. Control over the situation 6. Awareness 7. Expectations on report resolution 8. Victim blaming

Table 6: Codebook

Checking, nudging or scoring? Evaluating e-mail user security tools

Sarah Y. Zheng
UCL

Ingolf Becker
UCL

Abstract

Phishing e-mail threats are increasing in sophistication. Technical measures alone do not fully prevent users from falling for them and common e-mail interfaces provide little support for users to check an e-mail's legitimacy. We designed three e-mail user security tools to improve phishing detection within a common e-mail interface and provide a formative evaluation of the usability of these features: two psychological nudges to alert users of suspicious e-mails and a "check" button to enable users to verify an email's legitimacy. Professional e-mail users ($N = 27$) found the "suspicion score" nudge and "check" button the most useful. These alerted users of suspicious e-mails, without harming their productivity, and helped users assert trust in legitimate ones. The other nudge was too easily ignored or too disruptive to be effective. We also found that users arrive at erroneous judgements due to differing interpretations of e-mail details, even though two-thirds of them completed cybersecurity training before. These findings show that usable and therefore effective e-mail user security tools can be developed by leveraging cues of legitimacy that augment existing user behaviour, instead of emphasising technical security training.

1 Introduction

E-mail has been one of the most pervasive forms of digital communication since the introduction of the internet. So much so, that the medium remains an attractive threat vector for adversaries to exploit [50]. Phishing e-mails, in which impersonated sources typically seek to gain money or sensitive data from a target recipient, have caused major security breaches, financial losses and psychological damage to unsuspecting users, making it a lucrative business for organised crime [15, 23, 33, 61].

Technical detection systems may capture the majority of phishing attacks, but do not fully prevent users from falling for them. As users are commonly regarded as the "last line of defence" [2], organisations invest in cybersecurity education for their employees and inform the public of potential scams. However, anti-phishing education and publicly available anti-phishing advice may not be as effective as hoped for [13, 38, 47, 57, 59].

An alternative way to help users disengage with suspicious e-mails is to enhance common e-mail interfaces to equip users with "just in time" decision-making tools. For instance, by nudging users to check sender information [49] or interactively showing the trustworthiness of URLs found in e-mails [54, 77]. Such developments showed promising results to decrease phishing susceptibility, but have been sparse and require further exploration [27].

Here, we provide an implementation-focused formative evaluation [71] of the usability of novel user-centric e-mail security concepts to help users detect suspicious e-mails in a common e-mail user interface (UI). First, we conceptualise (i) a "check" button to highlight indicators to help people assess both trustworthy and phishing e-mails, and (ii) a "collegiate phishing report" nudge and (iii) "suspicion score" nudge to make people aware of the possibility of phishing. We then examine how these security tool concepts affect users' e-mail processing behaviour, by collecting "think aloud" responses from professional e-mail users from one organisation (N=27) that processed e-mails in simulated Outlook e-mail interfaces without and then with the tools. Tool designs were updated following consistent feedback from at least five users over four iterations.

We find that the suspicion score nudge and final check button version were rated the most useful, and that people largely process the same pieces of e-mail information, but reason differently about them. This was surprising, as 18 of

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.
USENIX Symposium on Usable Privacy and Security (SOUPS) 2023.
August 6–8, 2023, Anaheim, CA, United States.

the participants recalled completing at least one mandatory cybersecurity training before and 19 have a technical study background. This implies that worse detection is not necessarily due to negligence of relevant security indicators [90], but a lack of consensus on how to interpret online information. For example, whether e-mails sent from free e-mail providers should be suspected in a professional context. This resulted in the following contributions:

1. We identify and discuss three fundamental trade-offs that can guide further development of usable e-mail security tools: (i) highlighting cues of desired (i.e., legitimate) vs. undesired (i.e., phishing) communication, (ii) enhancing users' existing behaviour vs. technical knowledge and (iii) not harming productivity for security.

2. We open-sourced our methods and data via GitHub and Open Science Framework (OSF) to encourage more studies on e-mail user security tools. This includes the simulated Outlook UI and two adapted e-mail sets to fit participants' organisational context to closely mimic an e-mail processing experience.

2 Related work

2.1 User-centric security interventions

People are thought to be bad at detecting phishing e-mails due to a lack of cybersecurity knowledge or awareness [5, 7, 38, 80, 83] and incautious e-mail processing behaviour [22, 31, 36, 44, 46, 76, 90]. The majority of interventions to improve human phishing detection thus focused on developing training and education programs [27]. Examples range from conventional education materials [6, 13, 16, 34, 36, 68, 82, 88] and serious games [9, 19, 28, 30, 40, 41, 69, 84], to phishing simulations [8, 18, 39, 40, 78]. While they increase user awareness of phishing threats, these programs require more frequent engagement than often is the case to stay effective in the long term [13, 59]. This waning effect may be due to the timing of educational programs before or after, but not during critical decision-making moments [27].

A different stream of user-centric security interventions aimed to make people aware of security-relevant information *during* decision-making. Earlier works used browser-based security warnings to prevent users from browsing suspicious domains [4, 26, 58, 67, 87]. Although such warnings are moderately helpful, they are prone to warning fatigue [4]. Alternatively, digital signatures may be used to add trust signals to e-mails [86], although criminals typically adopt them too once their use becomes widespread, as happened with SSL certificates [1].

Others have highlighted dubious domains [43] or e-mail sender information [49], but found limited detection improvements. This is likely due to users' misinterpretation of URLs [5]. A more promising approach provided users with interactive URL reports when they engaged with links found in e-mail messages [7, 54, 77]. By informing users about why a URL may be suspicious, these tools provide both educational and awareness-raising value. These works underline the potential of user-centric security interventions embedded in e-mail interfaces.

We expanded on this idea of aiding users during decision-making in an e-mail UI [27, 90]. Specifically, we used two under-explored concepts in e-mail security to design novel security features: 1. **enhancing users' confidence in trusting legitimate e-mails**, instead of following the predominant paradigm of enhancing phishing detection, and 2. **psychological nudges**, where slight changes in a UI improve decision-making, without forcing users to engage with those changes [72]. For instance, participants in a simulated e-commerce purchase displayed more secure behaviour when they received a notification that emphasised how they can cope with online shopping risks [73]. A study on phishing detection used a social alert in suspicious e-mails, saying that a high percentage of colleagues received the same e-mail [49]. Even though these works found small detection improvements, these findings suggest that nudging users with short and directly applicable information embedded in task systems can improve security behaviours. The potential of e-mail security nudges is discussed further in Franz et al. [27].

In line with the two concepts, we devised a "check button" to help users asses any e-mail's legitimacy, and two different nudges to alert users of suspicious e-mails. The first nudge contained a social cue, similar to Nicholson, Coventry, and Briggs [49], and also explained users what suspicious signs to look out for. The second nudge alerted users of potentially suspicious e-mails and showed recommended actions. With these features, we aimed to improve phishing detection by better supporting how users reason about e-mails while they are processing them.

2.2 Processing e-mail for communication versus security

E-mail processing has been described as comprised of "primary" and "secondary" tasks in the cybersecurity context [64]. The primary task refers to the main function of e-mail, i.e., communicating with others through digital means, which involves scanning, prioritising and responding to e-mail messages. The secondary task is the security check to decide if an e-mail is in fact legitimate and responding accordingly. On the one hand, we cannot reasonably expect users to focus on the secondary task [64]. On the other hand, even if we do, making the secondary task the main focus will inflate users' suspicions [53, 66, 70]. It is therefore a vital research challenge to design user-centric security interventions that augment users' secondary e-mail processing task, without harming their primary task.

The first step towards this goal is a deep understanding of

how users switch between the primary and secondary task in real-world contexts. Various studies have characterised aspects of how users process e-mails and the usability of adapted e-mail interfaces [11, 12, 21, 29, 45, 51, 79], but these provide limited or no insight into how users switch to the secondary task. The few qualitative works that do focus on how users reason about phishing find that both experts' and non-experts' e-mail processing involve understanding the e-mail context, finding surprising elements that lead to suspicion and acting on that suspicion [81, 83]. However, as these studies used phishing detection as the primary task [81] and relied on respondents who remembered a previously received phishing e-mail [83], it is unclear to what extent these results generalise to real-life contexts.

Thus, to engage users with the secondary task when they should suspect an e-mail, we need to understand how users change their reasoning from the primary to the secondary task. Then we can see how our proposed security features affect users' processing behaviour. Hence, we first analyse how users process e-mails without any security interventions, and then how our designs affect this behaviour.

3 Methods

Our formative evaluation [71] focuses on understanding what drives (un)usability of our novel e-mail user security tool concepts. To this end, we used qualitative methods to obtain an in-depth understanding of how our e-mail security tool designs affected users' e-mail processing behaviour and iterative design to make small short-term adjustments according to consistent user feedback. In this section we describe the study setup, the principled approach to our designs, our participants sample and analysis.

3.1 Participants

Twenty-seven participants performed the in-person e-mail processing task. They were recruited through e-mail invitations sent to staff at the researchers' institute. We only recruited staff from our institute, because (i) all staff were known to be experienced e-mail users, (ii) they could come to the session in-person, (iii) they would be familiar with the presented task context (e.g. e-mails from the same institute), (iv) they are likely to be used to the Outlook e-mail client, as their professional e-mails are processed through Outlook, and (v) they represent a working office population that relies substantially on e-mail communication. All were compensated with a £20 Amazon voucher. Their roles ranged from support staff to lecturers. The study was approved by our departmental Ethics Committee.

3.2 Task

To understand how users process e-mails, we asked participants to reason out loud while processing e-mails in simulated inboxes. They were told the study aimed to gather feedback on the usability of new e-mail interfaces and not security tools, to avoid biased responses [52]. We created a basic Outlook e-mail interface as shown in Figure 1, to which we added our security tool designs. Each participant had an in-person session of 45–60 minutes with the main researcher who sat down next to them.

After welcoming the participant and obtaining their informed consent, the researcher started an anonymous audio recording of the session. Participants first answered questions on their general e-mail use and were then instructed about the main task. They had to process e-mails as if they were professor Alex Carter in health informatics and talk through what they were doing and why. Participants were never told about phishing detection before or during the task. Only the security tool designs in the task could have prompted them to look out for phishing e-mails, as intended.

Each participant interacted with four different inboxes, one after another. They always started with the "control" inbox, i.e., without any new tools (Figure 1). The next three inboxes each contained one of the three security tools (see Section 3.3 and Figure 2) in random order. Each inbox contained eight or nine e-mails based on e-mails previously received by colleagues at the same institute to provide a familiar context (total $N_{legitimate} = 33$), and one or two phishing e-mails adapted from those previously received by academic institutes (total $N_{phishing} = 6$). Two phishing e-mails contained a malicious URL, purporting to be a Zoom meeting invite and Microsoft password reset. Two spearphishing e-mails seemed to come from professor colleagues requesting an urgent action. One phishing e-mail presented a fake paid mentor program, one "Nigerian prince"-style scam, and (see details on GitHub). Nearly all e-mails were made to directly address professor Alex Carter. Each e-mail could be replied to, forwarded, deleted, archived or moved to "junk". When participants wanted to reply to an e-mail, a text editor appeared through which they typed their reply and "sent" the message. We deemed these functionalities sufficient to simulate the experience of processing e-mails for human end users, as they cover the majority of user actions to process e-mails. This was confirmed by a participant's remark *"this feels like going back to work"* when they started the task.

To facilitate users with out-loud reasoning, the researcher asked participants to explain what they were looking at, to elaborate why they responded in certain ways to the e-mails, and, when participants explicitly mentioned that an e-mail looked legitimate or suspicious, why they thought so. The researcher also took written notes of any significant observations and asked if participants noticed any new feature when they did not interact with them during the first 2–3 minutes

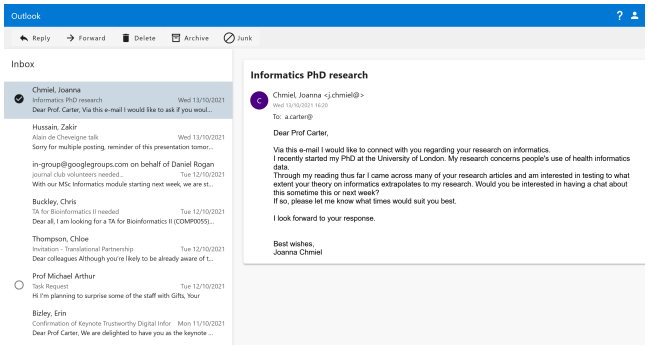


Figure 1: **Screenshot of an example e-mail in the "control" inbox.** The interface mimics the Outlook web client. Our security feature designs were added to this basic UI.

of viewing the inbox, to see why they had not (see study protocol on OSF). When doing so, we were careful not to mention any security concept, nor asking them to use the tool, to avoid biasing participants' processing behaviour. This way, any increased security awareness was merely the result of participants noticing the new security tool.

Every seven minutes, the next inbox automatically appeared, until participants saw all four inboxes. We kept the time spent on each inbox constant across participants to rule out the possibility that user engagement changed as a result of different times spent with a particular tool and ensure a study duration proportional to participants' compensation. We were aware that doing so traded off measuring detection accuracy for a reproducible qualitative method to evaluate usability. Efficacy will accordingly be described in terms of qualitative observation, not statistical comparison.

After completing the main task, participants gave feedback, voted which tool they found most useful and would use in real life, how many phishing e-mails they receive themselves, how much cybersecurity training they completed before and answered demographic questions (age, gender, education level, study background).

3.3 Rationale for tool designs

We designed our security features with two goals in mind: to help human users detect phishing e-mails and to assure users of the legitimacy of genuine e-mails. We took a principled approach. All designs had to be (i) user-centric, i.e., keeping humans "in the loop", (ii) accessible, i.e., easy to understand and use, and (iii) available at all times, which implies embedding new functionalities within the e-mail UI. The latter departs from conventional cybersecurity training, which may align with principles (i) and (ii), but not (iii). We defined three tool concepts: (i) "check" button, (ii) collegiate phishing report nudge, (iii) suspicion score nudge. We also kept the designs relatively small, in the form of an inbox add-on.

As part of our formative evaluation, we made small adjustments to the tool designs after at least five users gave us the same feedback. This resulted in four user-driven design iterations. The "check" button was updated three times, the "collegiate phishing report" nudge updated twice and the "suspicion score" nudge once. Figure 2 shows the final versions of the three tool designs, Appendix A depicts each iteration. Note that all updates were display-related changes and did not change key functionalities.

3.3.1 Check button

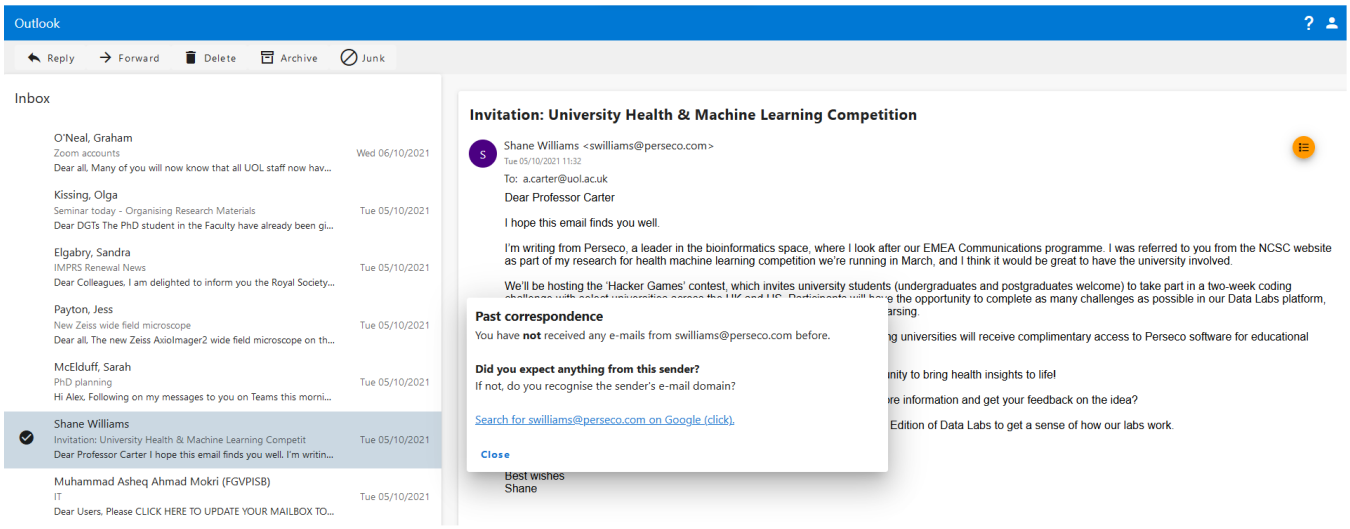
The first version of the "check" button sat in the task ribbon next to the "Junk" button. It aimed to provide users information on whether to trust a selected e-mail, based on common heuristics used by IT experts [81]. It could display overviews of (i) a dissection of the true URLs of any links found in the selected e-mail, since users often misinterpret URLs [5], (ii) the sender's name and e-mail address with short pieces of advice on what to do in case of mismatches in said details, and (iii) past e-mails received from the sender e-mail address. We expected these simple "checks" to help users when they are unsure if they could trust an e-mail by showing how URLs and sender details should be parsed and conjugated, as previous work implied that many users lack such reasoning [90].

Following consistent user observations, only the "past correspondence" check was kept in the last iteration. We placed the button closer to the e-mail sender details to which its functionality applied, following previous security feature design recommendations [74, 75], and simplified the button. If the user received e-mails from the sender's e-mail before, they are shown in a list with the date, time and subject line. If no past correspondence exists, the check information asks the user if they expected anything from the sender. If they did not, they are asked to double check the sender's e-mail domain. See Figure 2A and Appendix Figure 3.

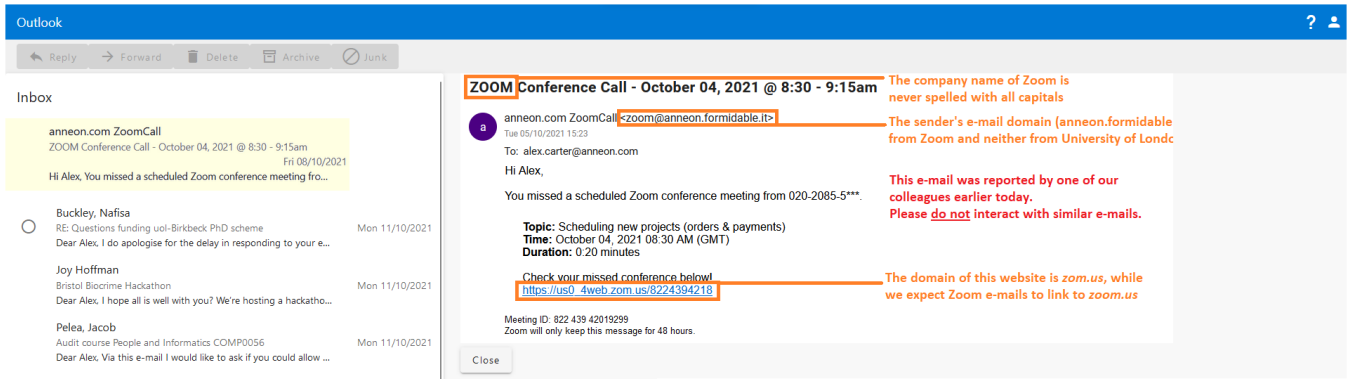
3.3.2 Nudge 1: Collegiate phishing report

The "collegiate phishing report" aimed to shift users' cognitive frame to investigating e-mail legitimacy [27], by using a socially oriented nudge that read "This e-mail was reported as suspicious today by one of our colleagues". The first version displayed this text in an orange warning banner between the Outlook task ribbon and e-mails display. When users clicked on it, a floating display appeared on top of the inbox with a screenshot of the phishing e-mail that was purportedly reported by a colleague, with annotations of all the suspicious cues in the e-mail that users had to look out for, and a general recommendation to not interact with similar e-mails. In the last iteration, the nudge looked like a new e-mail at the top of the e-mails list to increase user engagement. When users clicked on it, they saw the nudge text in the e-mail display with the fully annotated e-mail message and action recom-

A. Check button



B. Nudge 1: Collegiate phishing report



C. Nudge 2: Suspicion score

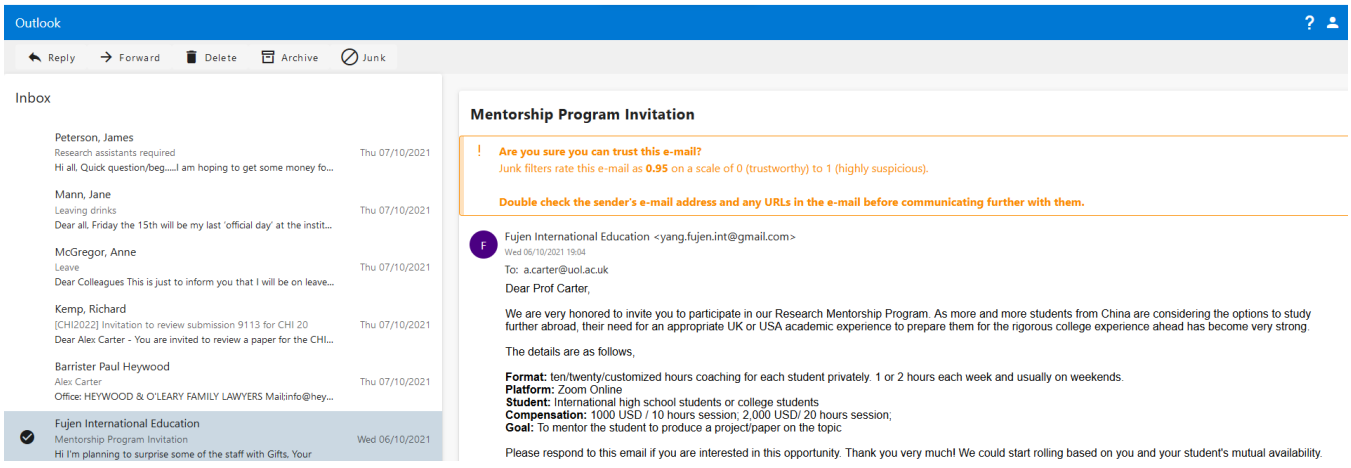


Figure 2: **Final designs of each security feature: A.** the past correspondence check button, **B.** collegiate phishing report nudge, **C.** the suspicion score nudge.

Iteration	Check button	Nudge 1: Collegiate phishing report	Nudge 2: Suspicion score
1 ($N = 8$)	The majority of users were unaware of the button until nudged towards it (after 2–3 minutes); users did not explore all sub menu items	Users tended not to click on the warning banner or got confused about which e-mail the warning is referring to	Users did not read all provided information, but found the orange colour positively alerting and useful
2 ($N = 7$)	Users remained unaware of the button until the researcher pointed it out, but also often did not see the benefit of the provided information.	Users did not like pop-up windows and often felt urged to close it right away	(design did not change)
3 ($N = 5$)	Users remained unaware of the button until the researcher pointed it out; the 'past correspondence' element was deemed useful	(design did not change)	(design did not change)
4 ($N = 7$)	Most users who noticed and started using the button found it very useful	More users skimmed over the warning content, some users found this and the suspicion score generally useful as they alerted them of suspicious e-mails	Users did not read all provided information, but found the orange colour positively alerting and useful; subtle text formatting edits did not lead to significantly more users applying the recommended actions

Table 1: Summary of user feedback on the security tool designs in each iteration

mendations. These annotations could not be removed and users could remove the nudge with the close button below it. See Figure 2B and Appendix Figure 4.

3.3.3 Nudge 2: Suspicion score

"Suspicion score" nudges were added to the phishing e-mails to prompt users to make a conscious effort to assess their legitimacy. This nudge was displayed in an orange warning banner between the task ribbon and e-mail display, and said "Are you sure you can trust this e-mail?". Below this line, it showed how suspicious the e-mail was on a scale from 0 to 1 and immediate action recommendations to nudge users' coping ability [73]. We chose to describe the score and scale to encourage users to think of potential false positives, e.g. when an e-mail has a lower suspicion score around 0.60, and thus take the recommended actions. Contrary to some inboxes that subtly add "(SPAM?)" in plain text to junk e-mail subject lines, we expected that our approach on e-mails in the main inbox would have a greater effect on users' vigilance and that explaining why the nudge was displayed would address user concerns over e-mails that could unwantedly be marked as junk. The amount of information was reduced and the text formatting changed slightly for the last iteration. See Figure 2C and Appendix Figure 5.

3.4 Thematic analysis

We used thematic analysis (TA) of a reflexive nature [17] to understand e-mail users' motives and considerations while they perform a typical e-mail processing task and evaluate how our tools affected these as a measure of usability. Given limited prior works on the qualitative relation between users' real-life e-mail processing behaviour and security, we interpreted the data inductively—without prior theories or hypotheses. After transcribing all session recordings, two researchers independently annotated one transcript, discussed the annotation codes and re-coded the same transcript to agree on the granularity of the coding approach. The first author, who conducted all participant sessions, annotated all remaining transcripts and freely added new codes to document all their user observations, in line with an interpretivist stance [17]. After completing all annotations, we iteratively extracted and refined themes through further discussions driven by the data. All transcripts and the full code book are available via OSF. We do not report inter-rater agreement scores, as they are inappropriate in reflexive TA [17].

4 Results

We performed an implementation-focused formative evaluation [71] of the usability of novel e-mail user security tool concepts based on highlighting legitimacy instead of phishing cues and nudging. We implemented the tool designs in simulated Outlook inboxes and let 27 professional e-mail users (mean age = 33.2 (SD = 7.2); 48% male; mean number of e-mail accounts = 4 (SD = 2.1); 19 with a Science, Technology, Engineering or Maths background) process e-mails in them while reasoning out loud. Their feedback was used as ongoing input for small short-term tool adjustments after at least five users provided similar feedback, resulting in four design iterations (Table 1).

Through thematic analysis, we gained a deep qualitative understanding of how our tools affected users' e-mail processing. Specifically, to understand why certain tools were (un)usable, it was key to understand how users reason without and then with our tools. A total of nine top-level codes and 86 secondary-level codes describe all user observations (see Appendix B and full code book on OSF). Based on these codes, we uncovered four overarching themes that capture how users reason about e-mails without our security tools (Section 4.1), and four themes that reflect how our security features affected them (Section 4.2).

4.1 Users' e-mail processing behaviour

The final codes naturally evolved to mirror users' primary and secondary e-mail processing behaviour. The "processing reasons" codes capture users' primary e-mail processing, and codes under "signals suspicious" and "signals non-suspicious" capture what users consider when they explicitly judge e-mails to be suspicious or non-suspicious (i.e., secondary e-mail processing). "Intended processing actions" reflect what users do with e-mails that do not raise any suspicion (i.e., primary processing) and "mitigation strategies" reflect how users assess and manage suspicious e-mails (i.e., secondary processing). Similar codes under "processing reasons" and "signals (non-)suspicious" suggest that users can arrive at opposing conclusions and perform different actions based on the same reasons (Section 4.1.3). Together with "prioritisation approach" and "prior experiences", these codes gave rise to the following four themes that describe how users generally reason about e-mails.

4.1.1 Content relevance

Most users first considered the relevance of the e-mail message content by judging the intent of the e-mail sender, before deciding what action to take. They either skimmed over subject lines, skimmed over the e-mail or read e-mails line by line right away. The importance of content relevance judgements is also reflected by the amount of "processing reasons" codes that relate to message contents ("high frequency", "important or urgent", "keep for reference", "not right audience or not personally targeted", "of personal interest", "outdated", "thread", "uninteresting or irrelevant"). These observations imply that users empathised with the e-mail task context.

The most common reason for users to find an e-mail suspicious was also based on e-mail message content. Out of all codes under "signals suspicious", "unexpected or funny content" has by far the most references. That is, most users seemed to assume legitimacy until they encountered e-mail content they perceived as odd. This accords with prior studies [52, 81], as well as psychological theory that people merely suspect things that are unlikely [24]. Further content-based reasons for users to be suspicious of an e-mail were "funny URLs", "requesting personal details", "urgent matter" and "fear appeal".

4.1.2 Relation to sender

Next, most users inferred their relationship with the perceived e-mail sender, by reading the sender's display name, a signature in the e-mail message and/or the actual sender's e-mail address. They made an assumption of how close they are to this sender according to the way the e-mail was written and the sender's e-mail domain. The "processing reasons" that reflect this theme are "unknown sender", "assume known or trusted sender", "from internal organisation", "automated e-mail", "newsletter". For example, user O112 assumed that the sender of a spear-phishing e-mail was indeed from the purported colleague professor: *"[. . .] a task request and it's from a professor and probably someone [who is] also a colleague of mine. And it's more personal because it starts with 'hi'."* This user assumed so, given the e-mail's informal writing style ("assume known or trusted sender"). To them, an "urgent request" may be reasonable to receive from a colleague and would not trigger suspicion.

Perceived closeness to e-mail senders was also the second most referenced factor that drove secondary processing. This is shown through "signals non-suspicious" codes "past correspondence", "internal e-mail", "trusted sender e-mail address" and "signals suspicious" codes "external sender", "non-professional sender e-mail", "no online info about sender organisation" and "unexpected sender or recipient name or e-mail address". An example of the latter, O11: *"This is a bit of a stranger and maybe someone genuinely called Olga Kissing. That is a bit suspect. So that depends on whether I actually knew that person. So I would just be suspicious from there."* The user reasoned that the sender's name sounded funny and therefore was untrustworthy, without reading the actual e-mail content. Only if they knew someone with that name, they might trust it.

4.1.3 Subjectivity in legitimacy perceptions

Users could have completely diverging assessments of the same e-mail, as the most attended to factors in e-mail processing described above are prone to subjective interpretation. For example, when users perceived an e-mail as not directed at them, we observed any of eight subsequent processing actions (see visualisations in Supplementary Materials on OSF). This aligns with prior work on user perceptions of "misdirected e-mail" [56]. We found bigger consensus among users that "unexpected or funny content" and "unexpected sender or recipient name or e-mail address", but not technical indicators, make an e-mail suspicious.

One scenario was that different users mentioned the same reason, but drew opposing conclusions for the very same e-mail. A prime example of this was in the case of a spear-

phishing e-mail sent from a GMail address purporting to be from a colleague professor. Out of all users who noticed this "unprofessional sender e-mail address", some users said it was junk straight away, whereas others responded without any suspicion. For example, user O19 commented: *"That sounds like a scam.[...] Because [of] the first bit. Also, it's from a GMail address."* First, they did not trust the e-mail because of the message content. They then noticed the sender's GMail domain, which further confirmed their suspicion.

Other users reasoned further about using private GMail addresses for work, e.g. user O14: *"If it's, like, a professor, same university. I would expect that communication would go in the same channel. So, like the University of London rather than a private e-mail. So I would maybe call that person and just, uh, ignore it, to be fair."* They would not expect a professional colleague to use a non-professional e-mail address to communicate with them. Their mitigation strategy would have been to call the sender to verify if they indeed sent the e-mail.

When viewing another e-mail from a GMail address, they described their suspicion of GMail accounts: *"And maybe they hate the Outlook interface by the university. [...] would not exclude it directly. That's the reason why I would look more on the content rather than, I mean, if it's like an e-mail, like an alpha numerical contact, like C H zero five, blah blah about to dot com, then I would think that not the right motivation is there."* They would consider both the e-mail message content and the sender's e-mail domain, but put more weight on the content. Similarly, user O26 appraised the e-mail content, but assumed that the same e-mail came from a known colleague, without mentioning any suspicion:

> *"Well, they've used a personal account, but they've signed it off as Professor Blackfield, and they work at the University of London [...] I would respond and say, sure, no problem. If it was someone I didn't know or the e-mail address was unfamiliar, I would probably ignore, delete. But in this instance [...] I presumed I know them. So I would say 'sure'."*

Another scenario was when users assessed different aspects of the same e-mail and drew opposing conclusions as a result. For example, O15 only looked at the message content of a phishing e-mail that indicated a missed Zoom conference and said: *"So this other e-mail is a Zoom conference call, but we missed it. If it is very important, [...] I would just mark it on my calendar to check this conference content or communicate with the people if necessary. But I would pin it if it is important."* They were not suspicious of the e-mail at all and overlooked the odd sender details and URL in the e-mail body. In contrast, user O27 first noticed unexpected sender details and marked the same e-mail as "junk": *"This e-mail address, Anneon.formidable, it is spam, so it's going to get junked. I do not do orders and payments. Somebody will tell me. I don't need an automated e-mail. So, that's junk."* It generally seemed that once any cue raised suspicion, users got rid of the e-mail as soon as possible or they started processing more information to substantiate their initial hunch—in line with findings from previous works [81, 83]. These examples show that within an e-mail, users assess different aspects, which leads to diverging legitimacy judgements.

4.1.4 User intents to UI functions

Our inbox simulation facilitated insights into how users translate their e-mail processing reasoning to how they interact with common inbox functionalities. We found that users had their own "mental models" of these functionalities. Strikingly, the vast majority of users deleted e-mails that they found suspicious, even though there was an option to mark e-mails as "Junk", e.g. O19: *"Junk it, delete it. Either way, get it out of the inbox. Like, I personally very rarely use junk to get rid of something."* This implies that most users do not distinguish between the type of "unwanted e-mails"—whether they thought the e-mails were uninteresting, irrelevant, or (potentially) malicious. They were usually treated the same way. It may thus not be practical for users to apply a different process to distinguish e-mails they found suspicious. User O13 even mentioned that they did not know that they could move e-mails to "Junk" themselves:

> *"**Researcher:** I also noticed that one of the e-mails that you thought was suspicious, you deleted it and you didn't say junk. Is that what you normally do as well?*
> ***User O13:** Yeah, I wouldn't necessarily say junk. That's not something we have, do we? [...] I never thought of to use that, possibly I'm ignorant. [...] I just, I never knew it existed, that we had a junk and we could put things in junk."*

Users also largely ignored the "Archive" button. E-mails that would be archived were usually deemed unnecessary and various users said they are unlikely to read archived e-mails ever again. As with suspicious e-mails, users may favour to delete anything irrelevant. A few users were concerned about e-mail storage and reasoned that deleting would therefore be better than archiving.

4.2 Effect of security features on users' e-mail processing

Thus far, we described users' overall reasoning while processing e-mails. Next, we examined how our security tools affected this processing behaviour. Through the iterative design process, we found that for tools to be usable, users value as little disruption to their primary task as possible and that simple changes such as the (symbolic) colour or placement of the tool matter more than content. As a result, the "suspicion

score" nudge received the most positive feedback in all iterations, with the "past correspondence" check from the "check" button as runner-up. In total, 9 participants found the "suspicion score" most useful, 7 participants the "check" button, 4 participants both the "suspicion score" and "check" button, 1 participant the "collegiate phishing report", 4 participants found all three designs most useful and 2 participants none of the tools.

All user interactions with and intra-task feedback on our tool designs were coded under "intervention feedback". When users adopted the given tool during the task and/or found it useful, it was coded as "positive". The other "intervention feedback" codes indicate points for improvement. Four themes emerged from these codes and "prior experiences", which together capture how users experienced the security features in relation to their e-mail processing. We summarised users' implicit (intra-task) and explicit (post-task) feedback in Table 1.

4.2.1 Usability of security information

Some "intervention feedback" pertained to the usability of information provided by the security features ("missing useful info", "functionality not clear", "too much information to process"). Fifteen users mentioned that the functionality of the "collegiate phishing report" nudge and/or the "check" button in the first three iterations was unclear, even though these tools were specifically designed to facilitate human interpretation of technical e-mail details. When users viewed the "check" button content in iterations 1–3, they often did not know how to interpret the provided information. For example, O12 in the first iteration:

> *"**Researcher:** When you see this, what are you thinking? [. . .]*
> ***User O12:** Whether there is malice or not. It's just not sufficiently. Well, it looks clunky [. . .] it would be much more useful to have something very clear saying 'this is safe, this is not safe' rather than giving me all this information."*

They were viewing a phishing e-mail and used the "check" button. After skimming over the provided check information, they pointed out that the information did not tell them whether to interpret the e-mail as suspicious or not. We expected users to be able to infer themselves whether they could trust an e-mail with the given details, but this did not seem the case. We considered displaying the information differently and adding more guidance in iteration 3, but they still found the amount of information too much, e.g. O16: *"Okay. Woah, so this confused me right away. So I just. Whatever. I just get out of here. Close. Because there's too much information. I don't understand anything. Lot of questions. A lot of, like, uh, sender details. And I don't know what they are."* We observed a similar negligence of technical information in the "suspicion score" nudge, which contained far less text and received most positive feedback. Most users did not read beyond the first line. Thus, providing users with more information to improve e-mail security seems to have no or even an adverse effect.

On the contrary, the "past correspondence" check was well received by all users. Some of them mentioned that they manually perform the same check in their own inbox, e.g. user O112: *"That will actually be useful, because I've found myself having multiple correspondence with the same person and I'll have to go and search for the name if I want to find something there. With that one, I think they're going to be much easier."* The provided functionality would save them time in real life, as they noted to regularly search for past e-mail correspondence with a given sender.

Other users indicated that the "past correspondence" check assured them of an e-mail's legitimacy, e.g. user O21: *"I just think it's fine, because there is a history of that e-mail, so it's fine."* This user noted that having exchanged e-mails before with a sender's e-mail address was a sign that the e-mail came from a trusted source. It did not necessarily matter to the user how many past e-mails were exchanged and about what. This mere fact indicated an established relation with the sender, which aligns with how users appraise their relation to a sender 4.1.2 and what has been described as "temporal embeddedness" as an indicator of trust [60]. Thus, users tended to ignore information that required more technical knowledge, but adopted information that augmented their existing e-mail processing behaviour.

4.2.2 Productivity versus security

In line with the previous theme, we found that users did not want to engage with features that interfered with their primary e-mail processing. This was most clearly observed with the "collegiate phishing report" nudge. Even if users did not read the provided content, we expected it to temporarily shift users' attention to the concept of e-mail legitimacy. In turn, this was expected to improve phishing detection. In most cases, however, users felt an urge to disregard or close the warning display as soon as possible when they saw it. For example, when user O16 in iteration 1 saw the warning nudge between the task ribbon and e-mails, they said *"'This e-mail was reported as suspicious by one of your colleagues.' Uh, did I do that? I use the cross."* They did not understand why they saw the nudge and closed it, without any further exploration.

In an attempt to increase user engagement with the nudge, iterations 2 and 3 showed the warning in a modal display in the newly loaded inbox. This was often experienced as highly disruptive. Yet, even in the last iteration when it was displayed as a highlighted e-mail that users had to click on themselves, most users only took a quick glance and closed it, e.g. user O14: *"So in this case, you have suspicion, my colleagues, with more information, uh, you missed a scheduled Zoom for blah, blah, blah. [. . .] Okay. Uh, I don't know. How can I get*

out of here?"

Some users who did read all the details in the "collegiate phishing report" nudge seemed to become more discerning throughout the task. Where they did not pay attention to the sender's e-mail address in previous inboxes, we found that they interpreted the sender e-mail address and explicitly mentioned whether a given e-mail was legitimate more often than before. It is unclear, however, for how long this effect would last. Together, these observations suggest that to improve users' secondary processing, we need to provide micro doses of information so not to harm users' primary processing.

4.2.3 User concerns on false positives

The "suspicion score" nudge aimed to alert users and enable better handling of suspicious e-mails. We found that users felt alerted and that most of them did not blindly delete or "junk" the given e-mail based on the warning. They often read the e-mail contents more carefully before deciding what action to take, which relates back to the content relevance theme in Section 4.1.1. While the majority of our users rated the "suspicion score" as the most useful feature, a few users expressed concerns around the possibility of a false positive warning. One of the phishing e-mails pretended to come from a Chinese company. When it contained the "suspicion score" nudge, user O211, of Chinese descent, mentioned that the warning was probably placed there due to algorithmic bias, as there is an allegedly large number of Chinese e-mail scams. They subsequently judged the e-mail as legitimate and ignored the recommended checks in the warning nudge that prompted users to double check the e-mail sender and links:

> *"**User O211:** This is Fujen International Education. This woman was. She is from China. [...] But I know the thing is, because I'm Chinese, I know lots of Chinese e-mails are flagged as not trustworthy, but this is just personal, so I'm going to forward it to my assistant instead of using this strong filter rating to decide [...] and we all know about algorithmic bias, do we?"*

With the increased attention for diversity and inclusion, users may grow especially sensitive to potential biases in automated decision systems. While the above e-mail was in fact a real phishing scam, it is important to take such user concerns into account when training detection systems and giving users security advice.

4.2.4 Ignorance toward security features

While the two nudges were ignored less, the "check" button remained untouched by 15 users until the researcher asked after three minutes if they saw the new button. Some users had seen it, but did not feel the need to explore it, e.g. user O114: *"[...] maybe I had seen it but I didn't really look at it and I didn't know what it was."* We found this surprising, as the button was located right next to the "Junk" button in the task ribbon in iterations 1–3 and next to the sender details in a different colour in iteration 4 to place it closer to the applicable e-mail content.

User O23 remarked *"Sometimes in the e-mails, you just go to, like, autopilot and you just... Yeah, I didn't even notice that."* This implies that new functionalities in e-mail UIs are easily missed, as users routinely process e-mails without thinking too much. Even after asking if users noticed the check button, not many consistently adopted its functionalities in the first three iterations. One explanation is that users did not find the information usable enough, as discussed in Section 4.2.1. This theme adds the possibility that users may have felt no need to use our features and preferred their own mitigation strategies when they suspected an e-mail (coded under "mitigation strategies"). These ranged from directly replying to the e-mail and judging by their response whether it was to be trusted, to asking colleagues and reporting it to IT. Together with Section 4.2.2, our findings suggest that users only adopt security features that align with their existing processing behaviour.

5 Discussion

As the sophistication of phishing attacks steadily grows [20], there is a pressing need to develop new e-mail security tools to protect users from falling for them. Here, we provide a formative evaluation of the usability of three e-mail user security tools based on the under-explored concepts of psychological nudges [27] and enhancing users' confidence in the legitimacy of genuine e-mails.

Through an iterative design process, we found that our past correspondence "check" button and "suspicion score" nudge help users detect phishing by affirming legitimate communication and alerting them of suspicious e-mails. These findings show the use of user-centric tool designs that enhance users' existing e-mail processing knowledge in a cost-effective way, instead of educating them about technicalities that many may find difficult to comprehend or easy to overlook. Future work could implement these designs into existing email clients and evaluate their effectiveness in-situ. Through these findings, we identified three usability versus security trade-offs, which we will discuss in light of developing usable e-mail security tools: highlighting legitimate vs. undesired communication (Section 5.1), supporting technical knowledge vs. existing behaviour (Section 5.2) and productivity vs. security (Section 5.3).

We also found that users largely process the same information found in e-mails (i.e., what the e-mail is about and whom the e-mail is from), but make judgement errors due to varying interpretations of those pieces of information—despite technical details explained in our tools. This is consistent with observations that users may notice surprising e-mail details,

but lack the knowledge to assess whether that information is suspicious [5, 35, 90]. Since all of our users are mandated to complete cybersecurity training (including phishing) every year, this was a surprising finding. It underlines the need for new approaches to improve detection [13, 38, 47, 57, 59].

5.1 Highlighting legitimate (desired) vs. undesired communication

The anti-phishing intervention literature typically focused on improving users' phishing detection ability by making them aware of suspicious cues in e-mails (see Section 2). Such "negative" framing is typically used in the wider usable security domain to alert users of potential security risks [37]. For most users, however, phishing e-mails likely comprise the minority of e-mails they receive. Our simulated inboxes therefore only contained 10–20% phishing. In these cases, users may rightfully assume that most e-mails are trustworthy, unless the prevalence of phishing e-mails is increased to a noticeable amount [66, 70]. This "prevalence paradox" limits the scope for user-centric anti-phishing tools within e-mail UIs, provided that user exposure to malicious e-mails remains relatively low. This implies that emphasising cues of legitimacy, such as with the past correspondence check, can maximise the scope to improve users' detection and confidence in e-mail legitimacy judgements.

Note that iterations 1–3 of the "check" tool explored multiple approaches to explain how to interpret technical sender and URL details: different buttons per functional content, simplifying language and information display, to no avail. As users were reluctant to process the provided details, we only kept the "past correspondence" check and then found positive user engagement with it. Although the "past correspondence" check could induce a false sense of security in cases of compromised or spoofed e-mail accounts, we believe the current findings provide a strong incentive to further explore e-mail user security tools that attend users to cues of legitimacy instead of phishing. For example, a next step could be to use language models to detect changes in the linguistic style of frequent senders to enhance the past correspondence check information and test it in an inbox that contains spoofed sender e-mails (of which we did not have examples).

In line with this paradigm and findings that many cybersecurity recommendations are unusable or vague [47, 57], our results further suggest that users would benefit from clearer organisational expectations on how to handle specific e-mail scenarios, whether legitimate or malicious. For example, tell users what to do with e-mails from free e-mail domains (e.g., Gmail). When users need to send or receive urgent requests, tell them what conventions to follow: e.g. to confirm with the person by phone or via a different channel than e-mail. This approach requires building a strong normative working culture that covers handling both legitimate and malicious e-mail communication. We would frame this contrast as "desired" versus "undesired" e-mail communication, instead of "phishing" versus "legitimate", as users may still have ill-defined ideas of what constitutes a phishing e-mail [20, 25].

5.2 Supporting (technical) knowledge vs. existing behaviour

Next, in support of Wash, Nthala, and Rader [83], we found that our most usable designs supported users' existing behaviour, whereas technical security-related content was ignored. Even within the "suspicion score" nudge, most users were sufficiently alerted by the mere presence of the nudge and did not read the recommended mitigation or suspicion score explanation. This aligns with recent quantitative findings from a large-scale study that more detailed warnings are not more effective than simple warnings [42], which highlights the power of our qualitative methods with a simulated environment. In the same vein, parsed URL and sender details shown with the original "check" button were also largely ignored. Users did not understand the utility of the provided information, even though the tool explained how to interpret the provided details. Users had a low tolerance for security information while processing e-mails.

These observations suggest that primary task interfaces may not be suitable for teaching users about e-mail security, which provides a further reason to refrain from using phishing simulations for "teachable moments" [42, 48, 65, 78]. Especially since all of our users are obliged to complete security and anti-phishing training and most of them still did not understand the provided security content, teaching users to accurately assess security information may be a task in vain. We therefore argue for an approach that leverages existing user behaviour, as was the case with the "past correspondence check" button and simple nudges like the "suspicion score".

Our finding that users were suspicious of e-mails when they perceived unexpected or "funny" content supports growing evidence that users pay most attention to e-mail content relevance, but not technical security indicators [32, 35, 52, 81, 83] such as URLs [90]. To then build on the idea of developing e-mail security tools that support existing user behaviour [83], another promising direction may be to categorise e-mails by their intent and show users nudges on e-mails with undesired discrepancies between intents and sender data. For example, when the e-mail message describes an "urgent request" and the sender is external to the user's organisation.

5.3 Balancing productivity vs. security

The last trade-off we found is how to balance users' productivity with security behaviour. The perception that security may be a (necessary) burden on users is a recurring theme [10, 14, 64] and we agree that security should not harm users' productivity. We also believe that well-designed security tools can

help users with minimal impact on their productivity, by leveraging how users actually reason about e-mails as described in Section 4.1, instead of explicitly trying to change users' (insecure) e-mail processing routines as suggested by [31]. The positive user engagement with the "past correspondence check" button exemplifies this. It relied on a heuristic that an e-mail can be trusted if there has been past correspondence with the sender's e-mail address. If there had not been any past correspondence, the button showed recommended actions to check the sender's legitimacy. This implicitly accords with our finding that users appraise their relation to the sender when viewing an e-mail. In contrast, the "collegiate phishing report" nudge was ineffective, as it was unclear to users how it related to the e-mails they wanted to process.

Another prominent observation is that most users never used the "Junk" button and several were confused about its functionality. This fits with findings that users' lack of understanding security concepts is associated with low or erroneous adoption of security tools [3, 63, 85]. Moreover, users did not distinguish between suspected phishing and generally unwanted e-mails, and just deleted both types. This possibly was the easiest action for them to understand and quickest to perform. We therefore expect that even with additional training, most users will refrain from using reporting functionalities, as these do not seem to align with how most users process e-mails and would affect their productivity. Even though the idea of personalised junk e-mail filtering systems has been suggested before [55, 89], security features that rely on machine learning models trained with users' processing behaviours (e.g., updating spam filters when users move e-mails to the junk folder [62]) may thus be unreliable. It may even be more beneficial to remove the "Junk" and "Report as phishing" buttons and "Junk" folder altogether. Taking into consideration recent studies [12, 51, 86], it is recommended that future research incorporates our suggested e-mail tool designs and assesses their efficacy in real-world settings. Furthermore, considering the themes and trade-offs highlighted in our research, there is ample opportunity to undertake more foundational UI design research pertaining to phishing.

5.4 Limitations

We did not test our features on live users or over longer time periods, as our formative evaluation aimed to first eliminate designs that provide low usability. In doing so, we tried to get as close to a realistic e-mail processing setting as possible by situating the study at an office desk in a regular office space, modelled the simulated inboxes after the institutional Outlook interface, participants participated at a time of their preference, and adapted e-mails received by colleagues at the participants' institute. Although the e-mails were unfamiliar to our participants, we did not expect this to fundamentally change how they reason about e-mails in the task. Indeed, our results without security features align with other qualitative works on how users process e-mails [32, 81, 83] and several participants remarked that the task felt like they were back at work. Lastly, we could not test if certain designs were better for certain demographic groups. Still, we would not expect significant differences by type of user, as we found a strong consensus among participants that the "suspicion score" nudge and "past correspondence" check were most useful.

6 Conclusion

Phishing is a persistent threat to organisations worldwide. Technical security measures alone do not sufficiently prevent people from falling for them, as users get exposed to new, evermore sophisticated attacks. Here, we provide a formative evaluation of three novel e-mail security tool concepts to help users discern trustworthy from suspicious e-mails in simulated inboxes. We used qualitative methods to gain a deep understanding of how our security tools affected user behaviour and thus why certain designs were (un)usable. Our "check" button supported user confidence in the trustworthiness of legitimate e-mails and the "suspicion score" nudge was deemed most useful. These findings highlight the potential of intuitive cues of legitimacy to augment existing user behaviour, instead of emphasising technical security knowledge. Together, they provide guiding principles for further usable security tool developments. We also found that users infer the trustworthiness of e-mails from the same types of information, but that differing interpretations of that information lead to erroneous judgements. This was surprising, as most of our users completed cybersecurity training before and have a technical study background. Future interventions that highlight desired versus undesired e-mail communication norms may help create more consensus among users. We hope these findings pave the way for a new generation of user-centric security tools to curb the risks of phishing threats.

7 Acknowledgements

We would like to thank the anonymous reviewers, Carlos Rombaldo Jr, Neil Amhis, Gerard Buckley and Nadine Michaelides for their valuable feedback on earlier versions of the manuscript. Sarah Zheng is supported by the UCL Dawes Centre for Future Crime, and Ingolf Becker is supported by the Engineering and Physical Sciences Research Council (grant number EP/W032368/1).

References

[1] Josh Aas et al. Let's encrypt: an automated certificate authority to encrypt the entire web. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, CCS '19, pages 2473–2487,

London, United Kingdom. Association for Computing Machinery, 2019. ISBN: 9781450367479. DOI: 10.1145/3319535.3363192.

[2] Jemal Abawajy. User preference of cyber security awareness delivery methods. *Behaviour & Information Technology*, 33(3):237–248, 2014. DOI: 10.1080/0144929X.2012.708787.

[3] Ruba Abu-Salma, M. Angela Sasse, Joseph Bonneau, Anastasia Danilova, Alena Naiakshina, and Matthew Smith. Obstacles to the adoption of secure communication tools. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 137–153, 2017. DOI: 10.1109/SP.2017.65.

[4] Devdatta Akhawe and Adrienne Porter Felt. Alice in warningland: a Large-Scale field study of browser security warning effectiveness. In *22nd USENIX Security Symposium (USENIX Security '13)*, pages 257–272, Washington, D.C. USENIX Association, 2013.

[5] Sara Albakry, Kami Vaniea, and Maria K. Wolters. What is this URL's Destination? Empirical Evaluation of Users' URL Reading. In *Conference on Human Factors in Computing Systems - Proceedings*. Association for Computing Machinery, 2020. ISBN: 9781450367080. DOI: 10.1145/3313831.3376168.

[6] Lukas Aldag, Peter Mayer, Mattia Mossano, Reyhan Duezguen, Bettina Lofthouse, Tatiana Landesberger, Melanie Volkamer, and Benjamin Berens. An investigation of phishing awareness and education over time: when and how to best remind users. In *Sixteenth Symposium on Usable Privacy and Security (SOUPS 2020)*, 2020.

[7] Kholoud Althobaiti, Nicole Meng, and Kami Vaniea. I don't need an expert! Making URL phishing features human comprehensible. *Conference on Human Factors in Computing Systems - Proceedings*, 2021. DOI: 10.1145/3411764.3445574.

[8] Aurélien Baillon, Jeroen De Bruin, Aysil Emirmahmutoglu, Evelien Van De Veer, and Bram Van Dijk. Informing, simulating experience, or both: A field experiment on phishing risks. *PLoS ONE*, 14(12), 2019. ISSN: 19326203. DOI: 10.1371/journal.pone.0224216.

[9] Malak Baslyman and Sonia Chiasson. Smells phishy? an educational game about online phishing scams. In *2016 APWG Symposium on Electronic Crime Research (eCrime)*, pages 1–11, 2016. DOI: 10.1109/ECRIME.2016.7487946.

[10] Adam Beautement, Ingolf Becker, Simon Parkin, Kat Krol, and M. Angela Sasse. Productive Security: A Scalable Methodology for Analysing Employee Security Behaviours. In Twelfth Symposium on Usable Privacy and Security (SOUPS 2016), Denver, CO. USENIX Association, 2016.

[11] Victoria Bellotti, Nicolas Ducheneaut, Mark Howard, Ian Smith, and R. Grinter. Quality versus quantity: e-mail-centric task management and its relation with overload. *Human-computer Interaction*, 20, 2005. DOI: 10.1207/s15327051hci2001&2_4.

[12] Frank Bentley, Josh Jacobson, Charlotte Sperling, Shiv Shankar, Chris Royer, and Ian McCarthy. Rethinking consumer email: the research process for yahoo mail 6. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI EA '20, pages 1–6, Honolulu, HI, USA. Association for Computing Machinery, 2020. ISBN: 9781450368193. DOI: 10.1145/3334480.3375224.

[13] Benjamin Berens, Kate Dimitrova, Mattia Mossano, and Melanie Volkamer. Phishing awareness and education – when to best remind? In *Workshop on Usable Security and Privacy (USEC)*, 2022.

[14] Denis Besnard and Budi Arief. Computer security impaired by legitimate users. *Computers & Security*, 23(3):253–264, 2004. ISSN: 0167-4048. DOI: 10.1016/j.cose.2003.09.002.

[15] Marzieh Bitaab et al. Scam pandemic: how attackers exploit public fear through phishing. In *2020 APWG Symposium on Electronic Crime Research (eCrime)*, pages 1–10, 2020. DOI: 10.1109/eCrime51433.2020.9493260.

[16] Jim Blythe, L. Camp, and Vaibhav Garg. In *Targeted Risk Communication for Computer Security*, pages 295–298, 2011. DOI: 10.1145/1943403.1943449.

[17] Virginia Braun and Victoria Clarke. One size fits all? what counts as quality practice in (reflexive) thematic analysis? *Qualitative Research in Psychology*, 18(3):328–352, 2021. DOI: 10.1080/14780887.2020.1769238.

[18] AJ Burns, M. Johnson, and Deanna Caputo. Spear phishing in a barrel: insights from a targeted phishing campaign. *Journal of Organizational Computing and Electronic Commerce*, 29:24–39, 2019. DOI: 10.1080/10919392.2019.1552745.

[19] Gamze Canova, Melanie Volkamer, Clemens Bergmann, and Benjamin Berens. Nophish app evaluation: lab and retention study. In *NDSS workshop on usable security (USEC 2015)*, 2015. DOI: 10.14722/usec.2015.23009.

[20] Fiona Carroll, John Adejobi, and Reza Montasari. How good are we at detecting a phishing attack? investigating the evolving phishing attack email and why it continues to successfully deceive society. *SN Computer Science*, 3, 2022. DOI: 10.1007/s42979-022-01069-1.

[21] Marta E. Cecchinato, Abigail Sellen, Milad Shokouhi, and Gavin Smyth. Finding email in a multi-account, multi-device world. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 1200–1210, San Jose, California, USA. Association for Computing Machinery, 2016. ISBN: 9781450333627. DOI: 10.1145/2858036.2858473.

[22] Hongliang Chen, Christopher E. Beaudoin, and Traci Hong. Securing online privacy: An empirical test on Internet scam victimization, online privacy concerns, and privacy protection behaviors. *Computers in Human Behavior*, 70:291–302, 2017. ISSN: 07475632. DOI: 10.1016/j.chb.2017.01.003.

[23] Xi Chen, Indranil Bose, Alvin Chung Man Leung, and Chenhui Guo. Assessing the severity of phishing attacks: a hybrid data mining approach. *Decision Support Systems*, 50(4):662–672, 2011. DOI: 10.1016/j.dss.2010.08.020.

[24] Morton Deutsch. Trust and suspicion. *Journal of Conflict Resolution*, 2(4):265–279, 1958. DOI: 10.1177/002200275800200401.

[25] Julie Downs, Mandy Lanyon, and Lorrie Cranor. Decision strategies and susceptibility to phishing. In *Proceedings of the second symposium on Usable privacy and security (SOUPS)*, pages 79–90, 2006. DOI: 10.1145/1143120.1143131.

[26] Serge Egelman, Lorrie Cranor, and Jason Hong. You've been warned: an empirical study of the effectiveness of web browser phishing warnings. In *SIGCHI Conference on Human Factors in Computing Systems (CHI)*, 2008. DOI: 10.1145/1357054.1357219.

[27] Anjuli Franz, Verena Zimmermann, Gregor Albrecht, Katrin Hartwig, Christian Reuter, Alexander Benlian, and Joachim Vogt. SoK: still plenty of phish in the sea — a taxonomy of User-Oriented phishing interventions and avenues for future research. In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*, pages 339–358. USENIX Association, 2021.

[28] C. J. Gokul, Sankalp Pandit, Sukanya Vaddepalli, Harshal Tupsamudre, Vijayanand Banahatti, and Sachin Lodha. Phishy - A serious game to train enterprise users on phishing awareness. In *CHI PLAY 2018 - Annual Symposium on Computer-Human Interaction in Play Companion Extended Abstracts*, pages 169–181, 2018. DOI: 10.1145/3270316.3273042.

[29] Catherine Grevet, David Choi, Debra Kumar, and Eric Gilbert. Overload is overloaded: email in the age of gmail. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, pages 793–802, Toronto, Ontario, Canada. Association for Computing Machinery, 2014. ISBN: 9781450324731. DOI: 10.1145/2556288.2557013.

[30] Matthew L. Hale, Rose F. Gamble, and Philip Gamble. Cyberphishing: a game-based platform for phishing awareness testing. In *2015 48th Hawaii International Conference on System Sciences*, pages 5260–5269, 2015. DOI: 10.1109/HICSS.2015.670.

[31] Jonas Hielscher, Annette Kluge, Uta Menges, and M. Angela Sasse. "taking out the trash": why security behavior change requires intentional forgetting. In *New Security Paradigms Workshop*, NSPW '21, pages 108–122, Virtual Event, USA. Association for Computing Machinery, 2021. ISBN: 9781450385732. DOI: 10.1145/3498891.3498902.

[32] Markus Jakobsson. The Human Factor in Phishing. *Privacy Security of Consumer Information*, 7:1–19, 2007.

[33] Jurjen Jansen and Rutger Leukfeldt. Coping with cybercrime victimization: an exploratory study into impact and change. *Journal of Qualitative Criminal Justice and Criminology*, 6(2):205–228, 2018.

[34] Jurjen Jansen and Paul van Schaik. The design and evaluation of a theory-based intervention to promote security behaviour against phishing. *International Journal of Human-Computer Studies*, 123:40–55, 2019. ISSN: 1071-5819. DOI: 10.1016/j.ijhcs.2018.10.004.

[35] Asangi Jayatilaka, Nalin Asanka Gamagedara Arachchilage, and Muhammad Ali Babar. Falling for phishing: an empirical investigation into people's email response behaviors, 2021. DOI: 10.48550/ARXIV.2108.04766.

[36] Matthew L. Jensen, Michael Dinger, Ryan T. Wright, and Jason Bennett Thatcher. Training to Mitigate Phishing Attacks Using Mindfulness Techniques. *Journal of Management Information Systems*, 34(2):597–626, 2017. ISSN: 1557928X. DOI: 10.1080/07421222.2017.1334499.

[37] Allen C Johnston and Merrill Warkentin. Fear appeals and information security behaviors: an empirical study. *MIS quarterly*:549–566, 2010. DOI: 10.2307/25750691.

[38] Iacovos Kirlappos and M. Angela Sasse. Security education against Phishing: A modest proposal for a Major Rethink. *IEEE Security and Privacy*, 10(2):24–32, 2012. DOI: 10.1109/MSP.2011.179.

[39] Ponnurangam Kumaraguru, Yong Rhee, Alessandro Acquisti, Lorrie Cranor, Jason Hong, and Elizabeth Nunge. Protecting people from phishing: the design and evaluation of an embedded training email system. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 905–914, 2007. DOI: 10.1145/1240624.1240760.

[40] Ponnurangam Kumaraguru, Steve Sheng, Alessandro Acquisti, Lorrie Faith Cranor, and Jason Hong. Lessons from a real world evaluation of anti-phishing training. *eCrime Researchers Summit*, 2008. DOI: 10.1109/ECRIME.2008.4696970.

[41] Ponnurangam Kumaraguru, Steve Sheng, Alessandro Acquisti, Lorrie Faith Cranor, and Jason Hong. Teaching johnny not to fall for phish. *ACM Transactions on Internet Technology*, 10(2), 2010. ISSN: 15335399. DOI: 10.1145/1754393.1754396.

[42] Daniele Lain, Kari Kostiainen, and Srdjan Čapkun. Phishing in organizations: findings from a large-scale and long-term study. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 842–859. IEEE, 2022. DOI: 10.1109/SP46214.2022.9833766.

[43] Eric Lin, Saul Greenberg, Eileah Trotter, David Ma, and John Aycock. Does domain highlighting help people identify phishing sites? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 2075–2084, Vancouver, BC, Canada. Association for Computing Machinery, 2011. ISBN: 9781450302289. DOI: 10.1145/1978942.1979244.

[44] Xin (Robert) Luo, Wei Zhang, Stephen Burd, and Alessandro Seazzu. Investigating phishing victimization with the heuristic–systematic model: a theoretical framework and an exploration. *Computers & Security*, 38:28–38, 2013. ISSN: 0167-4048. DOI: 10.1016/j.cose.2012.12.003.

[45] Gloria Mark, Shamsi T. Iqbal, Mary Czerwinski, Paul Johns, Akane Sano, and Yuliya Lutchyn. Email duration, batching and self-interruption: patterns of email use on productivity and stress. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 1717–1728, San Jose, California, USA. Association for Computing Machinery, 2016. ISBN: 9781450333627. DOI: 10.1145/2858036.2858262.

[46] Marijn Martens, Ralf De Wolf, and Lieven De Marez. Investigating and comparing the predictors of the intention towards taking security measures against malware, scams and cybercrime in general. *Computers in Human Behavior*, 92(November 2018):139–150, 2019. ISSN: 07475632. DOI: 10.1016/j.chb.2018.11.002.

[47] Mattia Mossano, Kami Vaniea, Lukas Aldag, Reyhan Düzgün, Peter Mayer, and Melanie Volkamer. Analysis of publicly available anti-phishing webpages: contradicting information, lack of concrete advice and very narrow attack vector. In *2020 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, pages 130–139, 2020. DOI: 10.1109/EuroSPW51379.2020.00026.

[48] Steven J. Murdoch and M. Angela Sasse. Should you phish your own employees? Bentham's Gaze. 2017. URL: https://www.benthamsgaze.org/?p=1756 (visited on 01/24/2023).

[49] James Nicholson, Lynne Coventry, and Pamela Briggs. Can we fight social engineering attacks by social means? assessing social salience as a means to improve phish detection. In *Proceedings of the 13th Symposium on Usable Privacy and Security (SOUPS 2017)*, pages 285–298. USENIX Association, 2017.

[50] Adam Oest, Penghui Zhang, Brad Wardman, Eric Nunes, Jakub Burgis, Ali Zand, Kurt Thomas, Adam Doupé, and Gail Joon Ahn. Sunrise to sunset: Analyzing the end-to-end life cycle and effectiveness of phishing attacks at scale. *Proceedings of the 29th USENIX Security Symposium*:361–377, 2020. DOI: 10.5555/3489212.3489233.

[51] Soya Park, Amy X. Zhang, Luke S. Murray, and David R. Karger. Opportunities for automating email processing: a need-finding study. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, Glasgow, Scotland Uk. Association for Computing Machinery, 2019. ISBN: 9781450359702. DOI: 10.1145/3290605.3300604.

[52] Kathryn Parsons, Marcus Butavicius, Malcolm Pattinson, Agata McCormac, Dragana Calic, and Cate Jerram. Do users focus on the correct cues to differentiate between phishing and genuine emails? In *ACIS 2015 Proceedings - 26th Australasian Conference on Information Systems*, 2015. ISBN: 9780646953373.

[53] Kathryn Parsons, Agata McCormac, Malcolm Pattinson, Marcus Butavicius, and Cate Jerram. Phishing for the truth: A scenario-based experiment of users' behavioural response to emails. *IFIP Advances in Information and Communication Technology*, 405:366–378, 2013. ISSN: 1868422X. DOI: 10.1007/978-3-642-39218-4_27.

[54] Justin Petelka, Yixin Zou, and Florian Schaub. Put your warning where your link is: Improving and evaluating email phishing warnings. *Conference on Human Factors in Computing Systems*, 2019. DOI: 10.1145/3290605.3300748.

[55] Vipul Ved Prakash and Adam O'Donnell. Fighting spam with reputation systems: user-submitted spam fingerprints. *Queue*, 3(9):36–41, 2005. ISSN: 1542-7730. DOI: 10.1145/1105664.1105677.

[56] Emilee Rader and Anjali Munasinghe. "wait, do i know this person?": understanding misdirected email. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 1–13, Glasgow, Scotland Uk. Association for Computing Machin-

ery, 2019. ISBN: 9781450359702. DOI: 10.1145/3290605.3300520.

[57] Elissa M. Redmiles, Noel Warford, Amritha Jayanti, Aravind Koneru, Sean Kross, Miraida Morales, Rock Stevens, and Michelle L. Mazurek. A comprehensive quality evaluation of security and privacy advice on the web. In *Proceedings of the 29th USENIX Security Symposium*, pages 89–108, 2020. ISBN: 9781939133175.

[58] Robert W. Reeder, Adrienne Porter Felt, Sunny Consolvo, Nathan Malkin, Christopher Thompson, and Serge Egelman. An experience sampling study of user reactions to browser warnings in the field. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, Montreal QC, Canada. Association for Computing Machinery, 2018. ISBN: 9781450356206. DOI: 10.1145/3173574.3174086.

[59] Benjamin Reinheimer, Lukas Aldag, Peter Mayer, Mattia Mossano, Reyhan Duezguen, Bettina Lofthouse, Tatiana von Landesberger, and Melanie Volkamer. An investigation of phishing awareness and education over time: When and how to best remind users. *Proceedings of the 16th Symposium on Usable Privacy and Security, SOUPS 2020*:259–284, 2020.

[60] Jens Riegelsberger, M. Angela Sasse, and John D. McCarthy. 53 Trust in Mediated Interactions. In *Oxford Handbook of Internet Psychology*. Oxford University Press, 2009. ISBN: 9780199561803. DOI: 10.1093/oxfordhb/9780199561803.013.0005.

[61] Fortune Jeff John Roberts. Exclusive: facebook and google were victims of $100m payment scam. 2017. URL: https://fortune.com/2017/04/27/facebook-google-rimasauskas/ (visited on 09/01/2022).

[62] Robert L Rounthwaite, Joshua T Goodman, David E Heckerman, John D Mehr, Nathan D Howell, Micah C Rupersburg, and Dean A Slawson. Feedback loop for spam prevention, 2007. US Patent 7,219,148.

[63] Scott Ruoti, Jeff Andersen, Scott Heidbrink, Mark O'Neill, Elham Vaziripour, Justin Wu, Daniel Zappala, and Kent Seamons. "we're on the same page": a usability study of secure email using pairs of novice users. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 4298–4308, San Jose, California, USA. Association for Computing Machinery, 2016. ISBN: 9781450333627. DOI: 10.1145/2858036.2858400.

[64] Angela Sasse, Sacha Brostoff, and D Weirich. Transforming the 'weakest link' — a human/computer interaction approach to usable and effective security. *BT Technology Journal*, 19, 2001. DOI: 10.1023/A:1011902718709.

[65] M. Angela Sasse and Steven J. Murdoch. Still treating users as the enemy: entrapment and the escalating nastiness of simulated phishing campaigns. Bentham's Gaze. 2021. URL: https://www.benthamsgaze.org/?p=3992 (visited on 01/24/2023).

[66] Ben D. Sawyer and Peter A. Hancock. Hacking the Human: The Prevalence Paradox in Cybersecurity. *Human Factors*, 60(5):597–609, 2018. ISSN: 15478181. DOI: 10.1177/0018720818780472.

[67] Stuart E. Schechter, Rachna Dhamija, Andy Ozment, and Ian Fischer. The emperor's new security indicators. In *2007 IEEE Symposium on Security and Privacy (SP '07)*, pages 51–65, 2007. DOI: 10.1109/SP.2007.35.

[68] Sebastian W. Schuetz, Paul Benjamin Lowry, Daniel A. Pienta, and Jason Bennett Thatcher. The effectiveness of abstract versus concrete fear appeals in information security. *Journal of Management Information Systems*, 37(3):723–757, 2020. DOI: 10.1080/07421222.2020.1790187.

[69] Mario Silic and Paul Benjamin Lowry. Using design-science based gamification to improve organizational security training and compliance. *Journal of Management Information Systems*, 37(1):129–161, 2020. DOI: 10.1080/07421222.2019.1705512.

[70] Kuldeep Singh, Palvi Aggarwal, Prashanth Rajivan, and Cleotilde Gonzalez. Training to Detect Phishing Emails: Effects of the Frequency of Experienced Phishing Emails. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 63(1):453–457, 2019. ISSN: 2169-5067. DOI: 10.1177/1071181319631355.

[71] Cheryl B Stetler, Marcia W Legro, Carolyn M Wallace, Candice Bowman, Marylou Guihan, Hildi Hagedorn, Barbara Kimmel, Nancy D Sharp, and Jeffrey L Smith. The role of formative evaluation in implementation research and the queri experience. *Journal of general internal medicine*, 21(Suppl 2):S1, 2006.

[72] R.H. Thaler and C.R. Sunstein. *Nudge: Improving Decisions about Health, Wealth, and Happiness*. Yale University Press, 2008. ISBN: 9780300146813.

[73] René van Bavel, Nuria Rodríguez-Priego, José Vila, and Pam Briggs. Using protection motivation theory in the design of nudges to improve online security behavior. *International Journal of Human Computer Studies*, 123:29–39, 2019. ISSN: 10959300. DOI: 10.1016/j.ijhcs.2018.11.003.

[74] Kami Vaniea, Lujo Bauer, Lorrie Faith Cranor, and Michael K. Reiter. Out of sight, out of mind: effects of displaying access-control information near the item it controls. In *2012 Tenth Annual International Conference on Privacy, Security and Trust*, pages 128–136, 2012. DOI: 10.1109/PST.2012.6297929.

[75] Kami Vaniea, Lujo Bauer, Lorrie Faith Cranor, and Michael K. Reiter. Studying access-control usability in the lab: lessons learned from four studies. In *Proceedings of the 2012 Workshop on Learning from Authoritative Security Experiment Results*, LASER '12, pages 31–40, Arlington, Virginia, USA. Association for Computing Machinery, 2012. ISBN: 9781450311953. DOI: 10.1145/2379616.2379621.

[76] Arun Vishwanath, Tejaswini Herath, Rui Chen, Jingguo Wang, and H. Raghav Rao. Why do people get phished? Testing individual differences in phishing vulnerability within an integrated, information processing model. *Decision Support Systems*, 51(3):576–586, 2011. ISSN: 01679236. DOI: 10.1016/j.dss.2011.03.002.

[77] Melanie Volkamer, Karen Renaud, Benjamin Reinheimer, and Alexandra Kunz. User experiences of TORPEDO: TOoltip-poweRed Phishing Email DetectiOn. *Computers and Security*, 71:100–113, 2017. ISSN: 01674048. DOI: 10.1016/j.cose.2017.02.004.

[78] Melanie Volkamer, Martina Angela Sasse, and Franziska Boehm. Analysing Simulated Phishing Campaigns for Staff. In *European Symposium on Research in Computer Security (ESORICS)*, pages 312–328. Springer, Cham, 2020. DOI: 10.1007/978-3-030-66504-3_19.

[79] Jaclyn Wainer, Laura Dabbish, and Robert Kraut. Should i open this email? inbox-level cues, curiosity and attention to email. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 3439–3448, Vancouver, BC, Canada. Association for Computing Machinery, 2011. ISBN: 9781450302289. DOI: 10.1145/1978942.1979456.

[80] Rick Wash. Folk models of home computer security. *ACM International Conference Proceeding Series*, 2010. DOI: 10.1145/1837110.1837125.

[81] Rick Wash. How Experts Detect Phishing Scam Emails. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW), 2020. ISSN: 25730142. DOI: 10.1145/3415231.

[82] Rick Wash and Molly M. Cooper. Who provides phishing training? facts, stories, and people like me. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, Montreal QC, Canada. Association for Computing Machinery, 2018. ISBN: 9781450356206. DOI: 10.1145/3173574.3174066.

[83] Rick Wash, Norbert Nthala, and Emilee Rader. Knowledge and capabilities that Non-Expert users bring to phishing detection. In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*, pages 377–396. USENIX Association, 2021. ISBN: 978-1-939133-25-0.

[84] Zikai Alex Wen, Zhiqiu Lin, Rowena Chen, and Erik Andersen. What.hack: engaging anti-phishing training through a role-playing phishing simulation game. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, Glasgow, Scotland Uk. Association for Computing Machinery, 2019. ISBN: 9781450359702. DOI: 10.1145/3290605.3300338.

[85] Alma Whitten and J. D. Tygar. Why johnny can't encrypt: a usability evaluation of pgp 5.0. In *Proceedings of the 8th Conference on USENIX Security Symposium*, SSYM'99, page 14, Washington, D.C. USENIX Association, 1999.

[86] Oliver Wiese, Joscha Lausch, Jakob Bode, and Volker Roth. Beware the downgrading of secure electronic mail. In *Proceedings of the 8th Workshop on Socio-Technical Aspects in Security and Trust*, STAST '18, San Juan, Puerto Rico. Association for Computing Machinery, 2020. ISBN: 9781450372855. DOI: 10.1145/3361331.3361332.

[87] Min Wu, Robert Miller, and Greg Little. Web wallet: preventing phishing attacks by revealing user intentions. In *Proceedings of the second symposium on Usable privacy and security (SOUPS)*, volume 149, pages 102–113, 2006. DOI: 10.1145/1143120.1143133.

[88] Weining Yang, Aiping Xiong, Jing Chen, Robert W. Proctor, and Ninghui Li. Use of phishing training to improve security warning compliance: evidence from a field experiment. In *HoTSoS: Proceedings of the Hot Topics in Science of Security: Symposium and Bootcamp*, pages 52–61. Association for Computing Machinery, 2017. DOI: 10.1145/3055305.3055310.

[89] Seongwook Youn and Dennis McLeod. Spam decisions on gray e-mail using personalized ontologies. In *Proceedings of the 2009 ACM Symposium on Applied Computing*, SAC '09, pages 1262–1266, Honolulu, Hawaii. Association for Computing Machinery, 2009. ISBN: 9781605581668. DOI: 10.1145/1529282.1529565.

[90] Sarah Zheng and Ingolf Becker. Presenting suspicious details in User-Facing e-mail headers does not improve phishing detection. In *Eighteenth Symposium on Usable Privacy and Security (SOUPS 2022)*, pages 253–271, Boston, MA. USENIX Association, 2022. ISBN: 978-1-939133-30-4.

A Screenshots of design iterations

Figures 3, 4 and 5 below show the design updates for each iteration for each of the three feature concepts. The updates mainly regarded positioning within the inbox interface, while keeping the contents the same. Only the "check" button design was updated more significantly for the last iteration compared to the initial design.

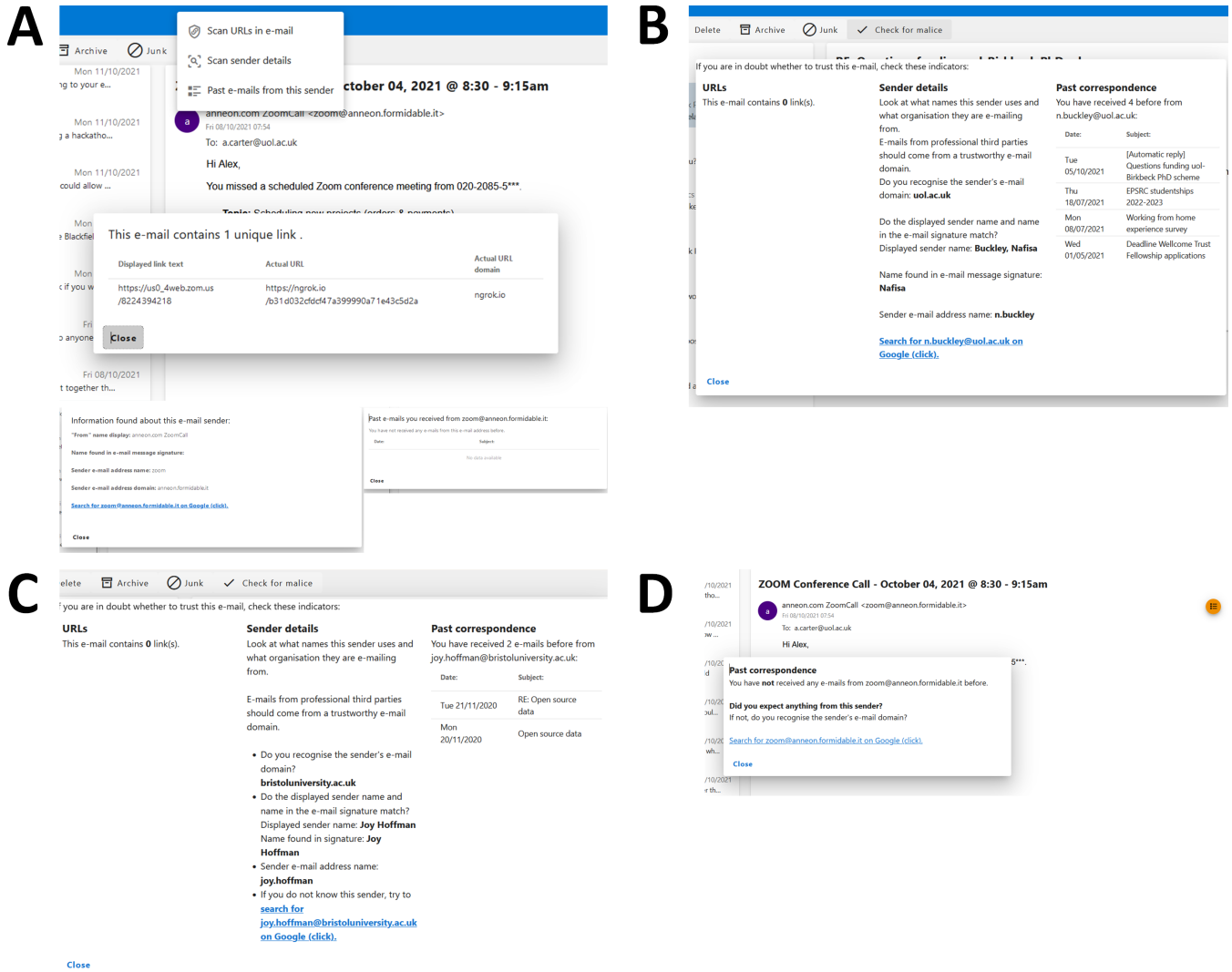


Figure 3: **"Check" button versions throughout the design iterations. A.** The first version of the button consisted of three sub menu items that appeared when a user hovered over the main button: (i) "Scan URLs in e-mail", which upon clicking would display an overlaid pop-up window with a table overview of all links found in the e-mail, the actual URL of the links and the actual URL domain; (ii) "Scan sender details", which displayed a list of sender name and e-mail address details as found in the user-facing e-mail header, as well as the e-mail body; (iii) "Past e-mails from this sender", which would display an overview of past e-mails received from the selected e-mail's sender e-mail address. **B.** After users pointed out the inefficiency of having three sub menu items to click in the first iteration's design, all three components from the first version were displayed at once when users clicked on the single "check" button. **C.** Users in the second iteration often found the displayed information too overwhelming (i.e., too much and/or too complicated) and tended not to read it. Hence, the amount of information was slightly reduced and displayed in a list for a more structured overview. **D.** In the last iteration, only the check for "past correspondence" was kept, since most users ignored or did not find the other technical information on URLs and sender details usable. We were aware of the potential false sense of security from this check in the case of spoofed e-mail addresses. Our task did not include spoofed sender phishing examples from the start, hence we used the last iteration to evaluate the usability of the concept of highlighting intuitive cues of legitimacy, rather than technical cues of malice. Also, to make the button more noticeable, it appeared as an orange icon next to the sender details.

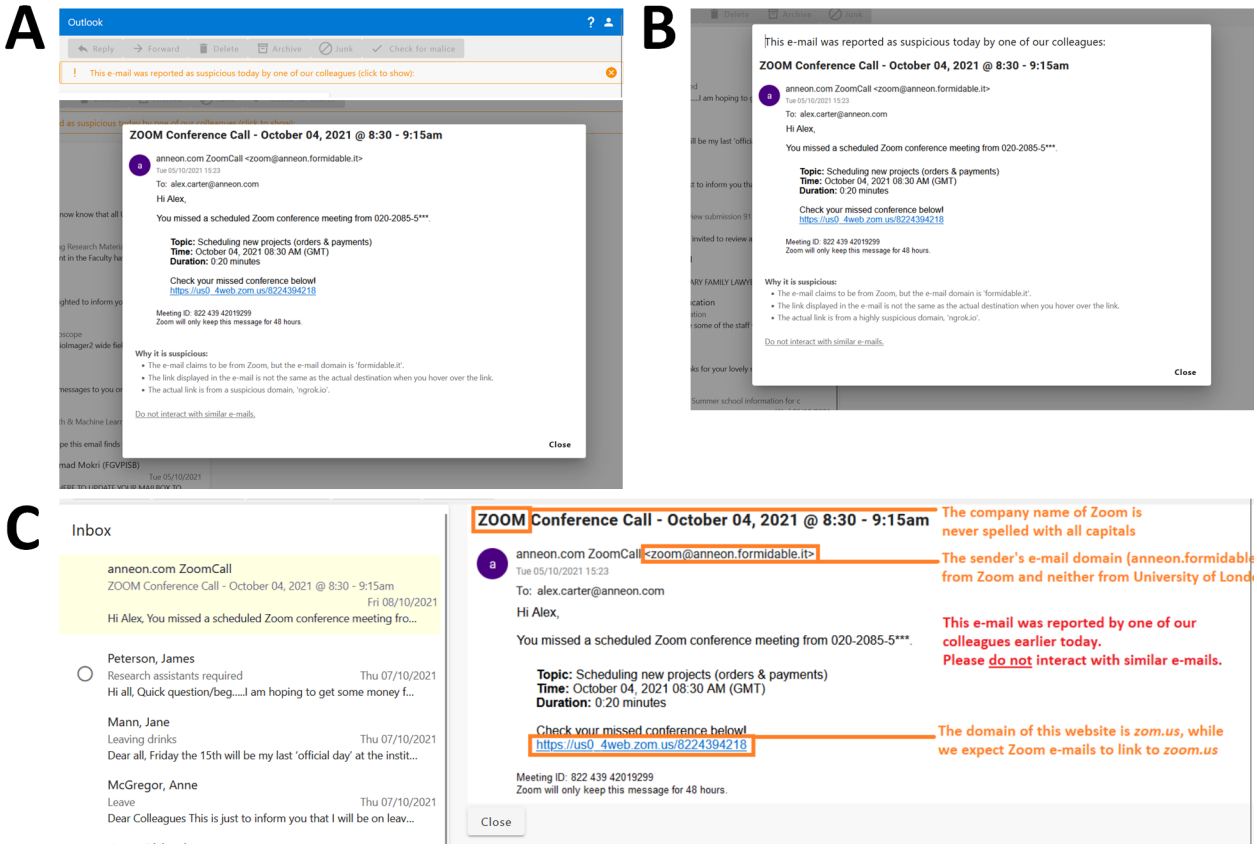


Figure 4: **"Collegiate phishing report" nudge versions throughout the design iterations A.** The first version was displayed as a warning banner between the task ribbon and the e-mails, which read "This e-mail was reported as suspicious today by one of our colleagues (click to show):" (left screenshot). If users clicked on it, a display on top of the inbox appeared with a phishing e-mail, a list with all reasons why it was malicious and that users should look out for similar e-mails (right screenshot). **B.** After many users either got confused, ignored or rapidly clicked away the warning banner in the first iteration, the updated version displayed the same display with the nudge text at the top when users loaded the new inbox. **C.** Feedback on the previous iteration indicated users' overall annoyance with the display. To make the nudge less disruptive, it was displayed as a highlighted e-mail at the top of the e-mails list. When users clicked on it, they saw the phishing e-mail with annotations in orange and the nudge text in red.

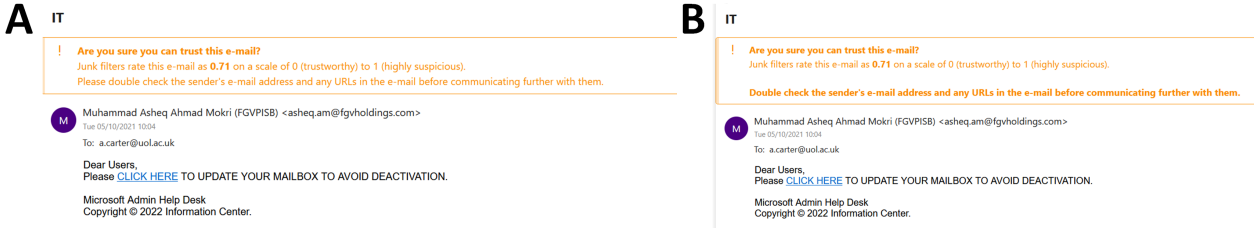


Figure 5: **"Suspicion score" versions throughout the design iterations. A.** Throughout the first three iterations, users were unanimously positive about the suspicion score design: an orange banner displayed on top of phishing e-mails with a score of 0.5 or higher. There were three lines of text. First a boldfaced line that warned users of whether they were sure they could trust the opened e-mail. Next, an explanation of why the warning is shown (high score on an automated suspicion scoring scale) and two recommended actions for the user (double checking links and sender details found in the e-mail). We showed the variability in suspicion scores for different suspicious e-mails to indirectly encourage users to think about the true legitimacy of a given e-mail. That is, to let them think of the possibility of misclassified "edge cases"—e-mails with relatively low suspicion socres (e.g. around .60). **B.** Users in previous iterations tended not to read the full warning text. Hence, the recommended actions were highlighted more by making them boldfaced and separating them with an extra line break for the last (fourth) iteration.

B Coding

Table 2 below shows all codes used to annotate users' reasoning and annotation frequency. The full code book including descriptions of each code is available via the OSF project page.

Top level	Secondary level	#
prioriti-sation approach	bottom-up	11
	quickly skim e-mails	42
	read whole e-mails	4
	senders-based	9
	top-down	40
	urgency	16
mitigation strategy	ask colleagues	3
	block sender	4
	blocked external sender content	1
	call sender	8
	check organisation via internet	13
	check past correspondence	2
	double check sender e-mail	14
	inspect linked page	3
	message actual internal sender	3
	not open attachment	1
	rely on antivirus to detect potential malice	2
	reply and evaluate response	8
	report to IT	1
	safe links	3
	train junk detection system	2
signals non-suspicious	formal e-mail signature	1
	internal e-mail	5
	IT Service Desk (ISD) checked	1
	looks important	3
	no warning sign	1
	not requesting sensitive information	2
	past correspondence	1
	proper written e-mail	3
	trusted sender e-mail address	15
	unclear reason	8
signals suspicious	external sender	26
	fear appeal	1
	funny URLs	21
	no online info about sender organisation	1
	non-professional sender e-mail address	41
	requesting personal details	10
	unclear reason	19
	unexpected or funny content	90
	unexpected sender or recipient name or e-mail	54
	urgent matter	3
	warning message	23
study feedback	design limitation	42
	unfamiliarity with context	14

Top level	Secondary level	#
processing reasons	assume known or trusted sender	46
	automated e-mail	8
	disseminate message	64
	from internal organisation	49
	high frequency	2
	important or urgent	37
	keep for reference	35
	meeting	113
	newsletter	4
	no action required	64
	no time for request	22
	not right audience or not personally targeted	36
	of personal interest	41
	outdated	12
	perform requested action or respond to query	84
	think about or research it before further action	41
	thread	21
	uninteresting or irrelevant	61
	unknown sender	8
intended process-ing actions	archive	66
	categorise in subfolder	14
	delete	157
	flag or pin	34
	forward	141
	junk	64
	leave in inbox	122
	reply	162
inter-vention feedback	functionality not clear	23
	missing useful info	11
	not applicable	1
	not sufficiently visible	2
	placement	3
	positive	15
	too much information to process	1
	unaware	17
	want to close nudge pop up asap	8
	warning could be false positive	4
prior experi-ences	experience academic context	1
	external sender warning	2
	known spam	8
	senders with unprofessional e-mail	1
	sensitised to security	2

Table 2: Final coding structure with reference frequency per code. Nine top-level codes and 86 secondary level codes were defined based on 27 session transcripts.

Understanding the Viability of Gmail's Origin Indicator for Identifying the Sender

Enze Liu, *UC San Diego*
Lu Sun, *UC San Diego*
Alex Bellon, *UC San Diego*
Grant Ho, *University of Chicago*
Geoffrey M. Voelker, *UC San Diego*
Stefan Savage, *UC San Diego*
Imani N. S. Munyaka, *UC San Diego*

Abstract

The current design of email authentication mechanisms has made it challenging for email providers to establish the authenticity of email messages with complicated provenance, such as in the case of forwarding or third-party sending services, where the purported sender of an email is different from the actual originator. Email service providers such as Gmail have tried to address this issue by deploying sender identity indicators (SIIs), which seek to raise users' awareness about where a message originated and encourage safe behavior from users. However, the success of such indicators depends heavily on user interpretation and behavior, and there exists no work that empirically investigates these aspects. In this work, we conducted an interactive survey (n=180) that examined user comprehension of and behavior changes prompted by Gmail's passive SII, the *'via'* indicator. Our quantitative analysis shows that although most participants (89%) noticed the indicator, it did not have a significant impact on whether users would adopt safe behaviors. Additionally, our qualitative analysis suggests that once prompted to consider why *'via'* is presented, the domain name displayed after 'via' heavily influenced participants' interpretation of the message *'via'* is communicating. Our work highlights the limitations of using passive indicators to assist users in making decisions about email messages with complicated provenance.

1 Introduction

Email is perhaps the longest-lived service in continuous use on the Internet and its precursor networks — dating back to at least Tomlinson's SNDMSG in 1971. As a result, email standards have not enjoyed the luxury of a careful design, but have instead accreted new mechanisms to shore up the legacy Simple Message Transfer Protocol (SMTP) against newfound problems. Chief among these problems has been email spoofing, whereby an adversary sends messages purporting to be from an address that does not, in fact, belong to them (e.g., for spam, phishing, etc.). To help mitigate such abuse, email protocol designers have added a range of out-of-band authentication protocols — SPF, DKIM and DMARC, among others — to help validate the identity of the sending organization (i.e., domain name) in an email message.

However, these mechanisms are hindered in practice because of modern Internet email borrowing heavily from the practices of mid-20th century business correspondence, including the notions of "carbon copies" (cc), message forwarding, and distribution lists.[1] In particular, both email forwarding and distribution lists require that messages be distributed by a third-party who is not the original sender — highly similar to spoofing. Thus, there are a range of legitimate scenarios where existing email authentication protocols will fail to validate the identity of the sender. To deal with this ambiguity, many email service providers (e.g., Google's Gmail and Microsoft's Outlook 365) choose to prioritize *deliverability* over possible security threats and will allow many such messages to reach user mailboxes [51].

This situation leaves individual users with the burden of distinguishing spoofed email messages from those messages that were merely ambiguously sourced. Moreover, the standard information displayed by a Mail User Agent (MUA) (e.g., To:, From:, Subject:, Date:, etc.) does not provide any indication that such a situation is even present, let alone provide sufficient evidence for making an informed decision. Spero and Biddle identify this issue as well, opining that "making the Mail-from (the true origin of the email) more visible would be beneficial, along with some information about the Mail-from domain" [74]. Gmail is one of the only two MUAs that attempts to inform their users of such situations, by providing a 'via' indicator in its user interface. Thus, a message from "alice@foo.com via bar.com" is intended to convey that the message claims to originate from `foo.com`, but was actually delivered by `bar.com`.

However, the utility of this indicator depends on the extent to which users understand its meaning, intuit its purpose, and are able to apply that understanding to then make informed

[1] The incorporation of these norms into email systems dates back at least to 1978, with Shoen's 1978 Mail client distributed with BSD Unix.

choices. In this paper, we examine the utility of Gmail's 'via' indicator to answer the following questions:

RQ1 How do users respond to the 'via' indicator?

RQ2 How would users react to the email when the 'via' indicator is present?

RQ3 What message is the indicator communicating to end users?

RQ4 What are users' perceptions of the relationship between the two domains shown by the indicator?

We answer these questions by surveying Prolific gig workers about the Gmail indicator. We replicated the Gmail interface, and asked participants to interact with a message from "alerts@chase.com" and answer follow-up questions about their experience. We employ a mixed-methods approach to our analysis to understand their interpretations of the 'via' indicator and identify how users respond to Gmail providing the indicator. We consider our results to be an upper bound baseline since gig workers are often more skilled in using various technologies.

Our results suggest that even with years of email experience, the 'via' indicator is not a factor in users' email decision-making process. Most of the participants (89%, n=120) that were shown the indicator remembered seeing it during the study. However, even in the case where the domain name displayed after 'via' (hence referred to as the 'via' domain) was `rlxaz.xyz`, most participants (85%, n=60) still believed Chase Bank or `chase.com` was the sender. Among these participants, 78% of them were "very confident" about their answers. Our results also suggest that the 'via' domain directly impacts users' interpretations of the indicator's purpose. In particular, users believed that the email they viewed was coming from `chase.com` through another part of the Chase Bank business when the 'via' domain was `chasesupport.com`.

These findings suggest that passive indicators that rely on user interpretation are likely to have limited success. In our study, once participants were asked to meditate on the purpose of 'via' from different perspectives, their interpretation evolved such that some participants completely changed their interpretation of the email. We suggest that future sender identity indicators be designed to communicate the necessary information users need without additional prompting. Ultimately, we make the following contributions:

- We provide an overview of how end users interpret the 'via' indicator and the factors that influence these beliefs.
- We present one of the first comparisons of user behavior in response to the indicator, a result that complements prior research on warning design and phishing susceptibility research.
- We identify challenges in communicating sender identity to technically experienced users and discuss how these barriers increase user risk.

2 Related Work

Our work falls under the domain of phishing prevention and email spoofing. We start by reviewing the prior literature on phishing prevention and then discuss relevant literature on email spoofing.

2.1 Phishing Prevention

Prior work on phishing prevention focuses on three main areas: (1) understanding users' phishing susceptibility and improving phishing training; (2) automatically detecting phishing attacks without user interaction; and (3) warning users about potential risks.

2.1.1 Phishing Susceptibility and Training

Because phishing exploits human mistakes rather than software vulnerabilities [98], researchers have investigated the reasons why users fall for phishing attacks and how to improve anti-phishing training and educational material. Prior work has found a variety of tactics that can make phishing email messages more persuasive [6, 9, 14, 30, 50, 57, 60, 79, 91], including having recognizable logos, targeting recipients' specific contexts, and using persuasive techniques, among others. Similarly, papers have identified a wide range of factors that affect users' susceptibility to phishing attacks [8, 9, 16–19, 29, 36, 50, 57, 59, 60, 71, 78, 79, 84, 86, 92], such as their age, personal traits, prior training, gender, and strategies they employ to detect phishing email messages. Leveraging these insights, other studies have focused on improving anti-phishing training [18, 19, 36]. This prior work includes exploring the efficacy of different training formats such as embedded training [11, 42, 45], teaching anti-phishing via games [45, 72, 87], and how the effectiveness of training varies in different contexts [37, 41, 43, 44, 46, 64, 66, 73, 85].

2.1.2 Automated Detection

Automated detection systems serve as the first line of defense by identifying attacks before users see them. The community has used a variety of algorithms to detect phishing email messages, websites, and URLs. These approaches range from commercial spam filters [63] to heuristics [13, 34, 40] and machine learning models [1, 20, 25, 75] proposed in academic work.

To detect attacks, these algorithms extract features from an email message, URL, and/or website and then apply a set of rules or machine learning model to identify phishing attacks. Prior work has explored a variety of different feature

sets [27, 54, 55, 83, 89, 94] and algorithms [38, 88, 99] to improve detection accuracy. Finally, simple approaches such as blocklists of IP addresses and accounts [7, 28, 53, 62] are also widely deployed in practice for phishing detection.

While beneficial, these detection systems face practical limitations. They can produce a large number of false positives when deployed at scale due to the high volume of benign email messages [61]. They also can be evaded by sophisticated adversaries [31]. As a result, automated phishing detection algorithms are often paired with phishing warnings to improve their effectiveness [45, 61].

2.1.3 Phishing Warnings and Indicators

Phishing warnings and indicators complement automated detection systems by alerting users of potential risks and supplying additional information to help users make informed decisions. Prior work has proposed different kinds of phishing warnings, including Passpet [96], dynamic security skins [15], SpoofGuard [77], Trustbar [33], social saliency nudges [58], active warning dialogs [10], and phishing warnings employed by browsers such as Chrome and Firefox [2]. Past research on these indicators has shown that passive indicators such as security toolbars are ineffective [22, 93], and active indicators that interrupt a user's current task are more useful in practice [22]. There also exists ongoing research that investigates the effectiveness of different anti-phishing support systems [68] as well as how to better design inclusive email security indicators [97].

Beyond these high-level warnings, other work has found that even subtle warning design choices can have a noticeable impact on the efficacy of phishing warnings and indicators [26]. In terms of ineffective warning design, prior work has found that user habituation to warnings [22] and failure to present information in a succinct and understandable fashion [16,93] lead to poor warning efficacy. While Lin et al. [49] report that using only domain highlighting as a browser warning does not provide strong protection against phishing, Volkamer et al. [80] found that combining domain highlighting with forced attention to a browser's address bar largely improves phishing detection. They also noted, in a separate study [81], the potential benefits of providing just-in-time and just-in-place tooltips, which follow-up studies [61] have confirmed. Zheng et al. [100] investigated the (in)effectiveness of presenting users with full email header details. Examining the use of multiple defenses, Yang et al. [95] discovered through a field experiment that combining phishing training and active phishing warnings can significantly reduce the click-through rate.

2.2 Email Spoofing

The original design of the Simple Mail Transfer Protocol (SMTP) lacked authentication, making email spoofing both possible and common [51]. Prior work examines a range of techniques attackers can use to successfully send spoofed email messages. Hu et al.'s [35] measurement study showed that many major mail providers delivered spoofed email to user inboxes without noticeable errors or warnings, and Chen et al. [12] demonstrated how attackers can compose multiple inconsistencies in different mail servers and clients to reliably send a spoofed email. More recently, several papers have explored how attackers can abuse email forwarding to send spoofed email messages. This work includes Shen et al.'s [70] large scale analysis on email spoofing attacks, Wang et al.'s [82] study on email spoofing opportunities introduced by Authenticated Receiver Chain (a standard for verifying servers that forward email), and Liu et al.'s [51] study on attacks enabled by email forwarding. However, all efforts mentioned above focused on the technical aspects of email spoofing. Our work is one of the first to examine the effectiveness of Gmail's 'via' indicator designed to mitigate such attacks.

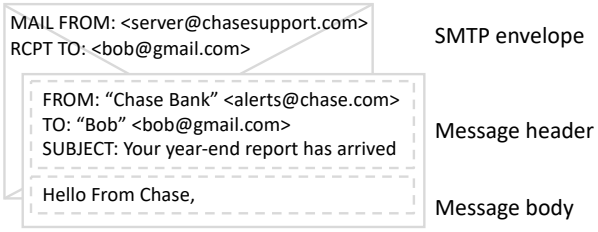


Figure 1: An example of SMTP headers, inspired by Figure 3 from Chen et al. [12].

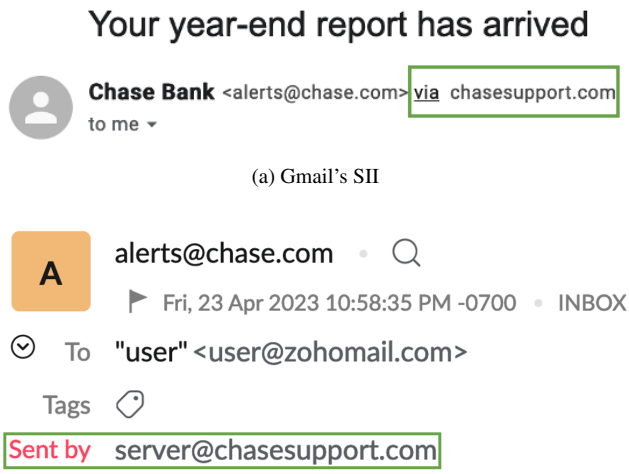
3 Background

In this section, we give a brief overview of SMTP (the protocol which governs the transmission of email) and provide background on sender identity indicators.

3.1 Simple Mail Transfer Protocol

Under the Simple Mail Transfer Protocol (SMTP), an email message includes two sets of headers that represent the sender(s) and recipient(s) of an email. Figure 1 shows an example message with both sets of headers. One set of headers, the SMTP envelope headers, consists of the MAIL FROM field and the RCPT TO field, and provides email servers with routing and delivery instructions. Specifically, the MAIL FROM field specifies the server that sent the email (`server@chasesupport.com`), and the RCPT TO field specifies the recipient of the email (`bob@gmail.com`).

The other set of headers, the SMTP message headers, includes the FROM and TO headers. This set of headers is used for user interface purposes only and does not affect email



(a) Gmail's SII

(b) Zoho's SII

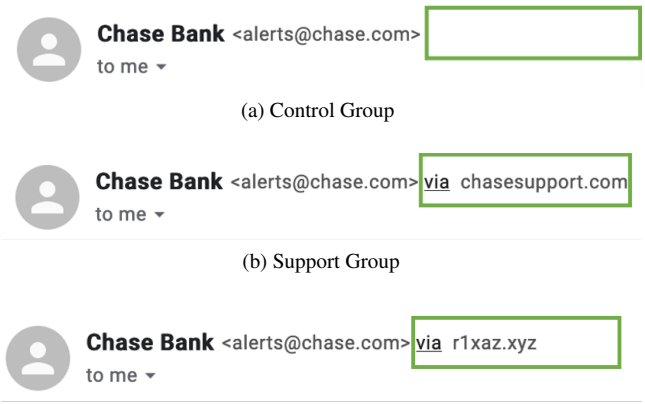
Figure 2: SIIs deployed by Gmail and Zoho with the SII highlighted.

routing [51]. Both the FROM and TO headers consist of a human-readable name and email address. In Figure 1, the FROM header consists of the human-readable name "Chase Bank" and the email address `alerts@chase.com`, and the TO header consists of the name "Bob" and email address `bob@gmail.com`.

3.2 Sender Identity Indicators

Under the SMTP protocol, the MAIL FROM and RCPT TO headers are opaque to users, and users only see the information in an email's FROM and TO headers. This design works well when the MAIL FROM and FROM headers share the same domain. In practice, however, the domains in an email's MAIL FROM and FROM headers do not always match. This mismatch occurs for a range of both benign and malicious reasons, including email forwarding (e.g., by mailing lists) and third-party sending email services, as well as email spoofing. To address the issue of header spoofing, the community has developed defensive protocols such as SPF and DMARC [21], where domains can provide information to recipients that allow them to validate if an email message truly originated from the domain, and that specify actions to take if such validation fails.

Unfortunately, due to limitations in these protocols and the lack of universal adoption, many recipient email servers cannot robustly authenticate all email messages. Moreover, in an effort to prioritize email deliverability [51], many domains often specify a permissive policy for recipients to follow if an email message fails to authenticate under a protocol like SPF or DMARC. As a result, major email providers such as Gmail and Microsoft Outlook often deliver email messages of unknown or potentially questionable authenticity.



(a) Control Group

(b) Support Group

(c) Random Group

Figure 3: Headers of the email shown to participants in each group.

To mitigate some of these issues, two email providers (Gmail and Zoho) have introduced UI modifications designed to provide additional information and awareness to users about an email's potential origins. We refer to these UI features as "sender identity indicators" (SIIs). Figure 2 shows an example of their SIIs with the indicator highlighted. In our work, we focus exclusively on Gmail's SII, the 'via' indicator, given Gmail's wide adoption [52]. Gmail uses 'via' to display the actual originator of an email message to recipients. For this message, the purported sender is `chase.com`, yet the actual originator is `chasesupport.com`. Once again, this mismatch can be due to using third-party sending services in a benign case, or email spoofing in a malicious case. Gmail cannot distinguish between the two cases, delegating the risk and leaving the decision up to the recipient (with the indicator as an aid).

4 Methodology

We use a between-subject study design to observe how the presence of the 'via' indicator impacts users' perception of the email sender when viewing an email in the Gmail interface. To answer these questions, we design a replica Gmail interface and survey participant groups under three different email conditions: "Control", "Support" and "Random". Participants in the *Control* group are presented with an email that has no 'via' indicator (Figure 3a). Participants in the *Support* group are presented with an email with the 'via' indicator (Figure 3b), which is followed by the `chasesupport.com` domain. Participants in the *Random* group are presented with an email with the 'via' indicator, which is followed by the `r1xaz.xyz` domain (Figure 3c). The Support group simulates a situation where the 'via' domain resembles the target domain, while the Random group simulates a situation where the 'via' domain is an unfamiliar domain. Participants were asked to log into a web interface modeled after Gmail, locate

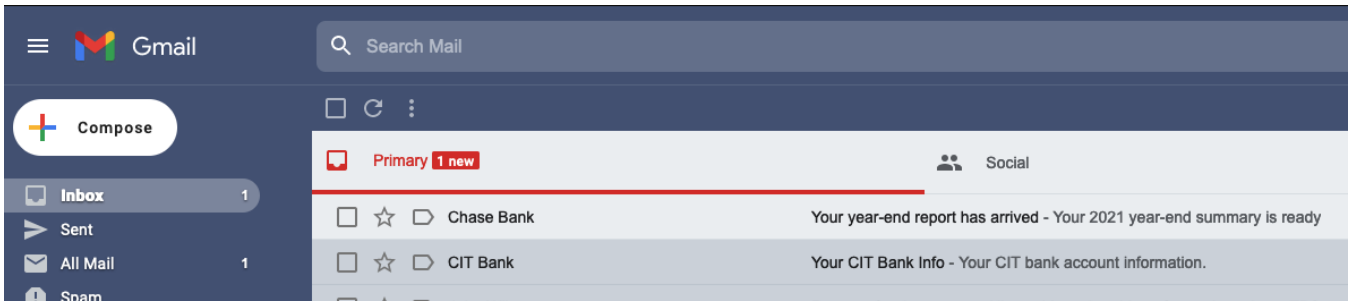


Figure 4: Email client landing page

and examine a specific email message, and answer questions about the email (Section 4.2). We randomly allocated participants to the three groups, which we will refer to below as Support, Random, and Control. We limit our scenarios to one email message and two 'via' domains to reduce the number of variables and focus specifically on understanding how people respond to 'via'. Below, we start by providing a brief description of our web-based email client and the email shown to the participants (Section 4.1), followed by a detailed description of our survey design and analysis methods.

4.1 Email Client

We built a web-based email client that is modeled after Gmail's web client. Our study focuses on Gmail because it is the most widely-used mail provider [52]. We decided to build a replica of the Gmail client instead of sending spoofed email messages to participants' real accounts so that we could easily track participants' interaction with the email in a controlled environment and avoid crossing ethical research boundaries.

Figure 4 shows the landing page of our web client. This page presents a list of email messages to the user, and we highlight the email that they need to review. We did not remove the Gmail brand name, as we seek to simulate users' experience with Gmail's web interface and increase ecological validity.

Upon clicking on the email that they are asked to review, participants are shown a page that displays the content of an email. This page mainly consists of two parts: email headers and email content. Figure 3 shows the email headers displayed to each of the survey groups. Users also have access to detailed header information that would be available in Gmail's web interface by clicking on the gray down-arrow button, also shown in Figure 3. Figure 5 shows the actual email content displayed. We take the content from a real email message sent by Chase that contains a link (the view my summary button) but change the link address to the main Google search page to prevent negatively impacting participants.

Lastly, we track if any of the buttons are clicked, if the link in the email is clicked, and when and how long users browsed the web interface.

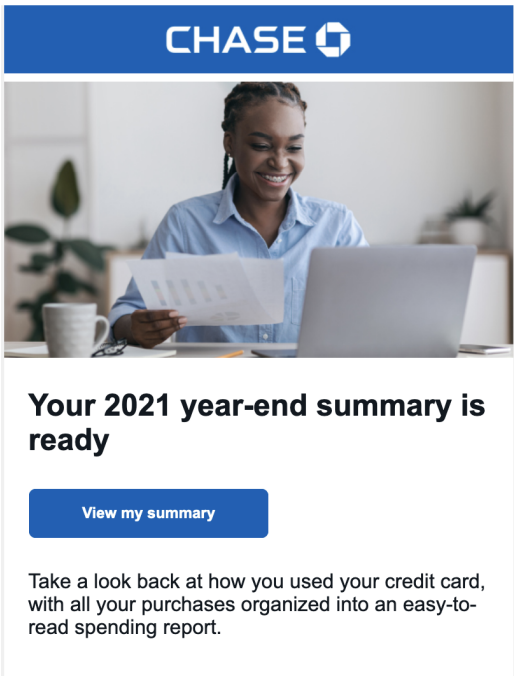


Figure 5: The email that participants needed to review.

4.2 Survey Protocol

We used Prolific to conduct our surveys. To avoid priming users for security, we framed the research as a study on the usability of the Gmail interface, including whether users are able to find an email and identify the sender of that email.

Since our study focuses on Gmail, we used a prescreening process to only include users with Gmail accounts. Specifically, we highlight in our survey description that participants must be Gmail users to enter the study. At the beginning of our survey, we also ask participants to confirm that they are indeed Gmail users and provide an option to exit the survey if they are not.

We give each user a unique link to our web-based email client (Section 4.1) after they pass prescreening. We embedded the link in the Qualtrics survey and instructed users to click on the link to access the email client in a separate tab,

Table 1: Demographics of survey participants

	Support *N* (%)	**Random** *N* (%)	**Control** *N* (%)
Age			
18-30	25 (42%)	28 (46%)	21 (35%)
31-50	30 (50%)	23 (39%)	30 (50%)
51-65	5 (8%)	7 (12%)	9 (15%)
Over 65	0 (0%)	2 (3%)	0 (0%)
Gender			
Female	22 (37%)	32 (53%)	28 (47%)
Male	36 (60%)	28 (47%)	32 (53%)
Non-binary	2 (3%)	0 (0%)	0 (0%)
Education			
No College degree	18 (31%)	22 (36%)	24 (40%)
2 year degree	5 (8%)	3 (5%)	2 (3%)
4 year degree	29 (48%)	24 (40%)	23 (38%)
Postgrad/Prof	8 (13%)	10 (17%)	11 (18%)
Prefer not to select	0 (0%)	1 (2%)	0 (0%)

and then return to Qualtrics to continue the study. In the survey, we started by asking users to imagine the email client was the actual Gmail web interface. Next, we instructed them to find, open, and read the email that was titled "Your year-end report has arrived". After this, we asked a series questions hosted with Qualtrics about the email (more details below). Participants had access to the email client throughout the study.[2]

First, we asked users to indicate the actions they would like to perform with the presented email by selecting from a list of available choices. This list of choices is adopted from prior work [18] and includes:

- Keep, save, or archive the email
- Click on the "View my summary" button in the email
- Forward the email to someone else
- Reply by email
- Contact the bank in other ways than email
- Delete the email
- Search a term in Google (please specify)
- Other (please specify)

We consider users who suggested that they would click on the link as having the potential to fall for phishing attacks, regardless of other actions they indicated.[3]

[2]Our survey questions, together with our implementation of the email client, can be found at https://github.com/ucsdsysnet/soups23-email-origin-indicator.

[3]While we did not have a follow-up phishing page that asked users to enter sensitive information, prior literature [41, 42, 71] has consistently suggested that 90% of the users who would click on the link would provide information on the phishing page.

Next, we asked users to answer three questions about the sender of the email: (1) the name of the person or entity that sent the email; (2) the email address of the person or entity that sent the email; and (3) how they decided the answer to the previous two questions. We also asked them to indicate their confidence level for questions (1) and (2) on a scale of 1 (not confident) to 5 (very confident).

We then moved on to ask users questions about the 'via' indicator. Specifically, we asked them to recall whether they saw the 'via' indicator during the study and whether they had encountered the 'via' indicator in the past before the study. We also asked them to indicate whether they understood what 'via' meant. For users who indicated that they knew the meaning of 'via', we followed up with a question asking them to explain what 'via' meant. For others who indicated that they did not know the meaning of 'via', we asked them to guess what information 'via' was trying to communicate. Lastly, for all users, we asked them to reflect on why Gmail chose to display the 'via' indicator.

Our last question probes the judgment made by users after having their attention directed to the 'via' indicator. We asked users to indicate whether they agreed that Chase.com used or instructed the 'via' domain (`rlxaz.xyz` or `chasesupport.com`) to send the email, and elaborate on their answer.

After answering the above questions and a demographic survey, users were debriefed about the true intention of this study and provided an option to have their data removed. We then thanked them for their participation and compensated them with $2.50 ($15/hr USD) for the 10 minute survey. We acquired approval from our institution's review board (IRB) before conducting the study.

4.3 Participants

Our sample size was informed by an a priori power analysis conducted with G Power [23] to determine the sample size needed for an effect size of .25 and alpha of .05 to test if one mean is significantly different among three groups. The results suggested a sample of 159 participants with 53 participants in each group for a power of .8. We received 180 unique responses across our three surveys, with 60 participants in each survey group. Most of our participants were between 31 and 50 years of age (46%), Male (53%), White (81%), and had a 4-year degree (42%). We compare the demographics of our participants, shown in Table 1, to the most recent US and UK Census data to evaluate how well they represent the US and UK populations. We saw that participants skewed toward younger age ranges than the US and UK populations, and higher educational attainment than the US population (but about the same as the UK population), meaning they were likely more familiar with computing concepts and usage.

In Table 2, we describe how familiar participants were with computers and phishing. The vast majority of participants

Table 2: Computer and email expertise demographics of survey participants

	Support Group *N* (%)	**Random Group** *N* (%)	**Control Group** *N* (%)
Computer Familiarity			
Work in or hold a degree in CS/IT	12 (20%)	6 (10%)	8 (13%)
Do not work in or hold a degree in CS/IT	48 (80%)	54 (90%)	52 (87%)
Computer Expertise			
Below or Somewhat Below Average	1 (2%)	1 (2%)	2 (3%)
Average	21 (35%)	22 (37%)	22 (37%)
Above or Somewhat Above Average	38 (63%)	37 (62%)	36 (60%)
Knowledge on Detecting Phishing			
Complete Novice	3 (5%)	1 (2%)	1 (2%)
Below Average	3 (5%)	4 (7%)	2 (3%)
Average	21 (35%)	31 (52%)	28 (47%)
Above Average	28 (47%)	20 (33%)	24 (40%)
Expert	5 (8%)	4 (7%)	5 (8%)
Years Spent Using Gmail			
Less than 4 years	13 (22%)	6 (10%)	11 (18%)
Greater than or equal to 4 years	47 (78%)	54 (90%)	49 (82%)

(86%) did not work in or have degrees in Computer Science, although 98% of participants across all surveys viewed their computer expertise as at or above average. Similarly, when asked about their skills detecting phishing, most participants claimed average or above average knowledge, with 8% on average claiming to be experts in detecting phishing. 85% of participants had been using Gmail for 4 or more years, meaning they were likely very familiar with its UI and accustomed to interacting with it. As a result, we present our findings as an upper bound on how average users will correctly absorb the information from Gmail's SII.

4.4 Analysis

Quantitative Analysis: We collected users' answers to multiple choice questions, multiple response (select all that apply) questions, open-ended question responses and their actions on our replica Gmail website. For our multiple response questions, we performed the test of proportions (z-test) to compare the responses following a prior study [39]. We cannot use a Chi-square test because the answers for our multiple response questions were not independently collected (i.e., a participant can choose multiple answers for each question). We used the Kruskal-Wallis test [90] and calibration curve [48] to compare confidence scores across groups.

We then use descriptive statistics to highlight the proportion of participants from each group that responded with specific answers. We present the proportion of participants that noticed 'via', selected specific behavior responses, and reported the proportion of participants with specific confidence scores. We use the statsmodels package in Python to conduct the analysis [69].

Qualitative Analysis: Two researchers on the team conducted iterative qualitative coding on the open-ended questions in the survey responses. (1) We asked participants to elaborate on the meaning of 'via' by asking "Please elaborate on what you think 'via' means". If participants reported that they did not understand the meaning of 'via', we asked them to guess: "What information do you think Gmail is trying to communicate by showing 'via' for this email? Please make your best guess and feel free to refer back to the email." (2) We asked participants to elaborate on the relationship between Gmail and 'via' by asking "Why do you think Gmail has chosen to display 'via' to users for certain emails?" (3) We asked participants to write down reasons "why they agree or disagree that Chase instructed the entity to send the email".

For each open-ended question, two researchers first conducted open coding to capture the major themes on 30% of the responses that were randomly selected. Then, the two researchers discussed and updated the codebook until an agreement about the themes was reached. We developed a codebook for each question to guide us in identifying the major themes for each condition. For example, participants were asked to explain why Gmail presented the 'via' in the email. One of the resulting codes for the Support group was "third party", which was used whenever a participant mentioned that Gmail provided the indicator to let them know the message was sent using a third party. Section A in the Appendix shows the code book and resulting themes.

After the training and codebook development, the two researchers coded all survey responses independently. After this initial coding, all codes reached acceptable inter-rater reliability (Cohen's Kappa above 0.7) [56]. The two researchers then talked through all instances where there was disagree-

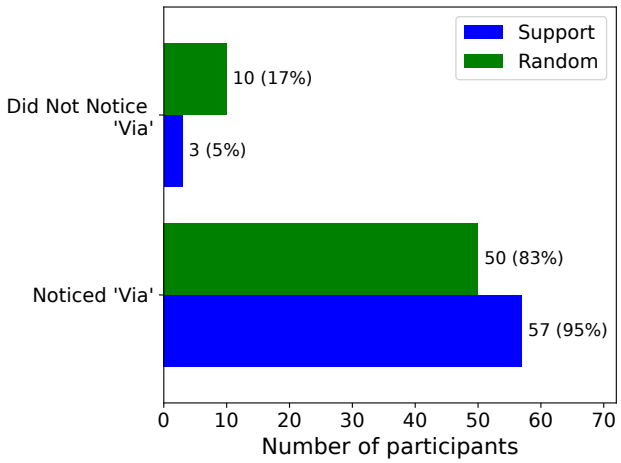


Figure 6: Number and percentage of participants that noticed 'via' in the study

ment and asked a third researcher to provide an opinion for judgment until a final decision was agreed upon.

5 Results

In this section, we present the qualitative and quantitative results of our study. We used this mixed-methods approach to understand participant behavior and indicator comprehension.

5.1 How do users respond to the presence of the 'via' indicator?

We explore the response of participants to the presence of the 'via' indicator by identifying (1) how many participants noticed 'via' when their attention is directed to the sender information section; (2) how many participants mentioned 'via' when determining the email sender; (3) how many participants checked the explanation of 'via' during the study.

The 'via' indicator was noticed by the majority of participants who were shown the indicator. Since security indicators are ineffective if they cannot capture users' attention [16], we asked participants if they noticed the 'via' indicator after they were asked to provide information about the email sender. Figure 6 shows that 89% of participants (n=120) noticed the 'via' indicator from the two groups that saw the 'via' indicator during the study. For the Support (n=60) and Random (n=60) groups respectively, 95% and 83% of the users in each group reported seeing the 'via' indicator during the study.[4]

While the notice rate is high, half of the participants believe they do not know the meaning of 'via'. After asking participants if they saw the 'via' indicator, we followed up by asking them if they knew the meaning of 'via'. Half of the participants (50%, n=120) reported not knowing the meaning of 'via' (22 in the Support group and 38 in the Random group). We hypothesize that this finding may be due to participant confidence and the limitations of 'via'. The indicator can only provide the origin domain for an email. It does not detect spoofing. Thus, instead of using 'via' as an aid to determine an email's origin, participants lean into their knowledge from prior experiences. Since they are confident about their ability to identify the email sender (the average confidence score is 4.61 and 4.70 out of 5.0 for the Support and Random group respectively), they might not care about the purpose or content of these indicators.

Given this low rate of understanding, we then examined the number of participants who clicked the indicator in our replica Gmail web browser, which provides an explanation of 'via'. Only 17 participants (4 in the Support group and 13 in the Random group) clicked the indicator while completing the study. We hypothesize that the low click-rate is mainly due to issues with indicator affordance — the indicator may not provide obvious visual cues that signal it can or should be clicked. Additionally, the fact that Prolific participants are motivated to complete the study quickly may have also contributed to the low click-rate.

Most participants did not mention 'via' when discussing the email sender. We asked participants to provide the email address of the sender, select how confident they were in their answer, and then discuss how they identified the information. Some users might not perceive the full difference between the email's true origin and its purported sender, but if the indicator works as intended, we expect experienced email users to acknowledge the via domain to some extent in their explanation, especially for the Random domain. Sadly, despite 89% of participants reportedly seeing 'via', and 50% of participants purportedly knowing the meaning of 'via', only 14% of participants mentioned the 'via' domain when explaining how they decided the sender of the email. Specifically, only eight participants (13%) in the Support group and nine participants (15%) in the Random group mentioned the 'via' domain to some extent in their answers. For example, P28 in the Support group specifically mentioned 'chase.com via chasesupport.com' and P52 in the Random group simply wrote *r1xaz.xyz* as their response. Lastly, only two participants, both from the Random group, raised concerns about identifying the email of the sender. For example, P40 in the Random group responded: *alerts@chase.com BUT there is a "via" thing after that that is new to me, and the explanation in the side window that pops up when I click on it about what "via" is, is not clear. If it weren't for that I'd be sure this came from Chase.com. But that "via" makes me wary.*

Additionally, despite most participants not mentioning the 'via' domain to some extent, the majority of participants were confident in their answer about the email sender when asked

[4] We note that prior work [24] has suggested that users can over-report their attention to security indicators. As such, our results represent the upper bound of the number of users who noticed the 'via' indicator.

Table 3: Confidence scores reported by participants when asked to provide the email address of the sender. The scale of 1 represents not confident and 5 represents very confident.

	Control n=60	**Support** n=60	**Random** n=60
5	50 (83%)	46 (76%)	46 (76%)
4	4 (7%)	8 (13%)	10 (17%)
3	1 (2%)	4 (7%)	4 (7%)
2	2 (3%)	1 (2%)	0
1	3 (5%)	1 (2%)	0

on a scale of 1 (not confident) to 5 (very confident): almost 80% of participants answered 5 (Table 3). After conducting a Kruskal-Wallis test, we found no statistically significant difference (p>.05) in confidence scores between each group. So while some users, like P40 in the Random group, were "wary", most participants were confident in their answers. This result suggests that, for most participants, 'via' is not a factor in identifying the origin of an email and does not lead to sender suspicion. The purpose of the indicator is to increase user awareness of email origin. If operating according to its purpose, more participants in the Random group would have confidently mentioned the 'via' domain (`r1xaz.xyz`) when discussing the email origin.

5.1.1 Calibration of Confidence

Following prior work [58, 59], we examine users' self-reported confidence against their actual performance in identifying the email address of the sender using a calibration curve [48], which is shown in Figure 7. The solid red line in Figure 7 represents perfect calibration, which is diagonal. When a user is perfectly calibrated, the probability of them mentioning 'via' in their answer is equal to their relative confidence in their answer (e.g., if a user is 80% confident in their answer, they would mention 'via' 80% of the time). Data points above the perfect calibration line correspond to users who are underconfident (e.g., if a user is 80% confident in their answer, they mention 'via' 90% of the time), while data points below the perfectly calibrated line correspond to users who are overconfident (e.g., if a user is 80% confident in their answer, they mention 'via' 70% of the time).

We derive the calibration curve for the Random group (orange dashed line) and the Support group (blue dotted line) by computing the rate of mentioning 'via' at each confidence level. We further convert the confidence level from 1 to 5 to a percentage scale (20% to 100%).

Overall, users are overconfident in both groups by a large margin. As past literature [65] has shown, confident users are less prone to change their online behavior and at a greater risk of being victimized. This result once again highlights that the 'via' indicator will not be effective in reducing unsafe behavior for users such as those in our study.

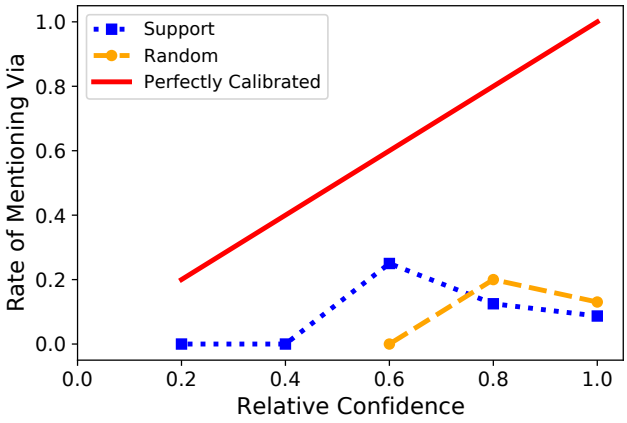


Figure 7: Calibration curve for identifying the email address of the sender.

5.2 How would users react to the email when the 'via' indicator is present?

To understand whether the 'via' indicator has an impact on participants' response to email messages that trigger 'via', we asked them what actions they would take after viewing the email shown in their group, as detailed in Section 4.2. Table 4 shows the number of participants from each group that selected the options provided. We compared the Random and Support group responses to the Control group responses using the test of proportions.

We did not observe statistically significant differences between the three groups (p>.05) for each action option. Most notably, over half the participants from each group selected that they would "Click on the view my summary button in the email". Also, none of the participants selected that they would "contact the bank in ways other than email", four participants selected they would "forward the email", and 38 (21%, n=180) participants selected that they would "delete the email", which are all actions Chase suggests people do if they receive a spoofed email [5]. This situation suggests that it is unlikely that the 'via' indicator encourages users to behave differently from when the indicator is not present.

5.3 How do users interpret the presence of the 'via' indicator?

Among the 120 participants that were shown the 'via' indicator, 47 participants (39%) marked that they knew its meaning. We examine if 'via' is effective at communicating the origin of the email to end users by asking all participants to explain or attempt to explain the purpose of the 'via' indicator.

Table 4: Number of participants from each group that selected the options provided

	Control n=60	**Support** n=60	**Random** n=60
Click button in email	39 (65%)	39 (65%)	35 (58%)
Archive the email	26 (43%)	30 (50%)	24 (40%)
Delete the email	13 (22%)	10 (17%)	15 (25%)
Other	2 (3%)	2 (3%)	5 (8%)
Reply by email	1 (2%)	2 (3%)	3 (5%)
Forward the email	0	1 (2%)	3 (5%)
Search Google	2 (3%)	1 (2%)	1 (2%)
Contact the bank	0	0	0

'via' means through. Many participants (21 in the Support group and 17 in the Random group) believed 'via' to mean through, as in this email was sent through a third party. For instance, P22 in the Support group wrote *"that the email came via an intermediary and not directly from chase.com*". In the Random group, P34 explained it as *"It's a return-path domain because the email was sent via a third party"*.

'via' indicates the sender. A significant amount of participants (18 in the Support group and 19 in the Random group) thought the 'via' domain indicated the true sender. For example, P21 in the Support group wrote *"That the e-mail was generated from a different domain name than the domain used in the actual e-mail sender @ address, I would assume like a mailer software that auto sent out the e-mails via a secondary support website."* and P6 in the Random group wrote *"That is the true origin of the email"*. P60 in the Random group took this explanation further and suggested that the email was forwarded, writing *"it could mean forwarded from i.e via but on reflection this is now likely a scam email phishing etc"*.

'via' indicates group association. When the target and 'via' domain include the bank brand name, participants associated the 'via' domain with the bank. Unique to the Support group, many participants (16) mention that 'via' means the email comes from an entity associated with Chase (e.g., Chase's support division). For example, P5 wrote that the email *"comes from a different department through the main company email"*. Additionally, two participants expanded on this idea, stating that 'via' indicated that an email was safe or authenticated. P25 from the Support group wrote *"That it's legitimately from Chase and not a scammer"*.

'via' encourages caution. When the 'via' domain does not include a brand name, participants explained that the presence of 'via' communicates a security risk. Many users (13) in the Random group explained that 'via' was being presented due to a scam or some other security risk that should be considered. While explaining the meaning of 'via', P59 in the Random group mentioned *"This [means] another website has been used to route the email. That's why I suspected it might be a phishing attempt"*. Additionally, P17 in the Random group wrote that the 'via' indicator *"possibly [means] that someone else is sending the email, looking at this more closely it appears like a scam"*.

This comparison suggests that the 'via' domain can influence users' interpretation of the 'via' indicator, especially when they try to guess the meaning. When the domain is shown as `chasesupport.com`, users are more willing to believe that this is related to Chase Bank, while the random domain triggered users' concerns about email safety.

5.4 What information do participants think Gmail is trying to communicate by showing 'via' in an email?

After participants explain what 'via' means, we then ask participants to reflect on why Gmail has chosen to display the 'via' indicator for the email they viewed in the study. In this section, we show how prompting participants to consider Gmail's perspective changes their interpretation.

In contrast to participants' explanation of 'via', security is one of the common reasons why many participants think Gmail has chosen to display 'via'. Of the 120 participants that saw 'via' in the study, 49 (41%) participants (27 from Support and 22 from Random) think Gmail chose to display 'via' to warn them of phishing or email legitimacy. For example, users explained that 'via' is displayed *"to make sure the email is not fraudulent"*[P18, Support]. Some users specifically mentioned security issues like phishing email or scam email: *"So people know the true email it came from cause it could actually be a scammer"*[P44, Random].

In addition to security, many participants believe Gmail uses 'via' to provide additional information about the sender. Many of the participants from both groups, 51 (43%) participants (24 from Support and 27 from Random) think 'via' is displayed to provide additional information. Some participants expressed this belief, writing that Gmail wants the user to know the email was outsourced to a third party or was not sent directly from Chase. Others explained it as Gmail wanting to provide additional transparency to the email. For example, P32 in the Support group elaborated that *"Gmail has chosen to display via to show where the email has come from if the recipient wants to check out the website."* However, some participants also connected this transparency to authenticating the sender — *"I think Gmail is adding it to certain emails to add authenticity"*[P9, Random].

5.5 What are users' perceptions of the relationship between the 'via' domain and `chase.com`?

After we asked users to think about what 'via' means and why Gmail chose to display it, we asked participants if they thought `chase.com` used the 'via' domain to send them the email and

explain their reasoning. The 'via' indicator only implies that the actual sender (which used the 'via' domain) of an email is different than the purported sender (with domain name chase.com), and the indicator was not intended to signal a relationship between the two domains.

Most participants from the Support (73%, n=60) and Random group (53%, n=60) believed that chase.com used the 'via' domain to send the email. Many participants (62%, n=120) believed that Chase Bank or chase.com used the 'via' domain to send the email because they believed the 'via' domain was the sending service, the domain chasesupport.com appears to be a part of the Chase business, or because this order of events matches their explanation of 'via'.

Some participants believed that chasesupport.com was a part of the Chase Bank business. Unique to the Support group, some participants (12) explicitly signaled the relationship between the domain chasesupport.com and Chase. For example, P18 described that *"even though both emails are from the same company, one division chase.com asked or used information from chasesupport.com."* This also includes participants who believe the email was initiated by Chase (5 participants) or that Google had verified the email (2 participants). Others (4 participants) in the group believed the two domains were associated, but that chasesupport.com instructed chase.com to send the email. An example of this perspective is from P21 in the Support group who wrote, *"The way that I think the 'via' works would, in my mind, mean that the chasesupport site auto-generated the e-mail and instructed the chase.com address to send the e-mail, not the other way around."* This result indicates the impact brand name has on user interpretations. We asked participants to explain the purpose of 'via' in their own words, from the perspective of Gmail, and then in relation to the target domain. In every scenario, multiple participants from the Support group viewed chasesupport.com as an authentic domain associated with Chase Bank.

Overall, only a small portion (6 in the Support group and 17 in the Random group) of participants were able to determine that chase.com did not use the 'via' domain to send the email and expressed some level of security concerns. Some of these users specifically mentioned the possibility of email being falsified and others raised some level of suspicion. P24 in the Support group correctly stated that *"The email was sent from chasesupport.com. chase.com didn't 'use' or 'instruct' anything. chasesupport.com sent the email"*. However, we note that even though this participant was able to correctly interpret the relationship between chase.com and the 'via' domain, they indicated that they would "Click on the view my summary button in the email" in the beginning of the study.

In fact, this apparent contradiction is not rare: after going through the questions, some participants realized what 'via' was communicating and expressed a new opinion of their previous actions. When discussing this question P8, in the Random group, wrote *"Had I seen the 'via' and the scary lookin' link, definitely would've just flagged this email, but it was sort of inconspicuous."* When asked how they would respond to the email, P8 selected: "Keep, save or archive the email"; "Click on the view my summary button in the email"; and "Delete the email". Thus, after taking time to reflect on 'via' from multiple perspectives, this participant was able to change their original decision. In fact, a non-negligible amount (5 in the Support group and 10 in the Random group) of participants expressed security issues after saying they would "Click on the view my summary button in the email" in response to the email earlier in the study. This result suggests that the indicator can be interpreted but is unlikely to nudge new behavior during the real-time decision-making process.

6 Limitations

The results of our survey are limited by the chosen scenario, the use of self-reported data, and participant demographics.

The Support, Random and Control group participants were all shown an email message that was supposedly from chase.com (which belongs to Chase Bank). As such, users' prior experience with Chase and prior exposure to email from Chase may have an impact on their responses. We also note that chase.com has a strong DMARC policy, and the example we show here is not representative of what might actually happen when chase.com is spoofed. However, research suggests that this type of spoofing with email forwarding can be done for other brands [51], thus we use chase.com to represent that possibility in the study. We chose a well-known brand name to investigate participant reactions when the target domain and 'via' domain include the brand name.

Next, while we strive to mimic users' real experience with Gmail, participants may act differently in our study compared to what they might do when using their personal email account. However, unlike other studies in this area, we focus on the differences in behavior selections due to indicator presence instead of focusing on one specific behavior under different email message conditions. We also recognize that users have access to the email client throughout the study, which may impact some of the self-reported results (e.g., "do you remember seeing 'via'?"). However, we believe users' response to this question does not negate their interpretation of the indicator's purpose. Design guidelines advise that warnings and indicators be noticeable and easy to interpret and, thus, should not require prior experience [47].

Lastly, our results may not generalize to the US and UK populations. Prolific users are more knowledgeable about security and have more confidence about that knowledge [76]. Due to this potential bias, we view our results as a reflection of how technically skilled individuals perceive the 'via' indicator.

7 Discussion

There are many individual challenges that, together, undermine a user's ability to effectively incorporate Gmail's "via" indicator into their decision making.

Among these are the general challenges associated with passive indicators (e.g., as highlighted in the context of phishing [22, 93]). Our work similarly documents that passive indicators are not able to prompt users to make safe decisions about email origin. This, in part, is because users often consider security as a secondary task [16] and rarely invest time and attention engaging in questions of security [32, 67]. In our work, we show that the presence of the 'via' indicator has little impact on whether or not users click on an embedded link. This result holds true even when we intentionally draw users' attention to the sender information section and even though the majority of users report that they noticed the 'via' indicator. In practice (i.e., without such prompting), it is likely that many users will overlook the 'via' indicator entirely: 'via' has a light gray color, is the same size as the rest of the header text, and is semantically vague without clearly conveying any notion of risk. The 'via' indicator could be made more noticeable if it were changed to a color that contrasts more with the surrounding text, and if the word used for the indicator were more related to its intended security purpose.

While this study focused on the 'via' indicator in the desktop version of Gmail, we also point out that the mobile Gmail app has no 'via' indicator at all. To get the same information provided by the 'via' on desktop, the mobile user has to open the collapsed box of sender details, and then click on the 'View security details' link to find the 'Mailed by' field. Using the phrase 'mailed by' instead of 'via' is an improvement, as it more accurately conveys the purpose of this indicator. However, the fact that this information is hidden behind two easy-to-miss interactions means that the chance of users finding this information is even lower.

Another major obstacle that hinders the success of 'via' is that it relies upon the ability of users to correctly interpret its meaning and the domain displayed after it. In our work, we show that the domain displayed after 'via' heavily influences users' interpretation of the indicator and their perception of the security risk. This once again highlights the potential issue of relying on users' computer knowledge. On one hand, past literature has consistently suggested that users cannot reliably determine the legitimacy of a domain [3, 4]. On the other hand, sometimes it is naturally a difficult undertaking to decide which domain names are connected with a specific organization [4].

A third issue, which is also common among security warnings, is the need for clearness in explanation. Participants in our study have indicated difficulty in comprehending the explanation of 'via' — if they were even able to find the explanation for the 'via' indicator in the first place. Indeed, upon carefully examining the current explanation of 'via', it does not convey the potential security risks in a straightforward way, and contains jargon that can be hard for users to understand. Given that the majority of Gmail users will not be familiar with DMARC policies or domain names, the current explanation does not fulfill its goal of explaining what the 'via' indicator means to the user. To improve both comprehension and safe behavior, the explanation should be updated with a more approachable explanation that highlights the potential security risks at the beginning, leaving the more technical details for the end so advanced users can still find it.

Last but not least, the current design introduces a new layer of complexity for users to determine spoofed email messages. Past papers have suggested that checking if the sender email address and organization are the same is a good approach to determine whether an email is legitimate [58, 100]. However, in the case of 'via', it is possible for the sender email and name to match, while also having a different domain as the 'via' domain. This situation makes it even more challenging for users to determine whether an email message is legitimate, as they are used to having to check that only two pieces of information match. This additional piece of information means users have to learn how to process more indicators, but once they learn how to do so, it can enable them to make safer decisions regarding which email messages to interact with. Gmail is one of the more secure email clients in this sense, as most clients do not display the 'via' information at all. These other clients without indicators prevent users from becoming too overwhelmed by information and warnings, but at the same time possibly expose them to greater risk of phishing or other unsafe situations. There is no an easy answer for this dilemma, since it involves a carefully balancing act of enabling users to make safe and informed decisions without succumbing to warning fatigue.

8 Conclusion

In this paper, we present a first analysis of the effectiveness and comprehensibility of Gmail's 'via' indicator. We conduct a survey to evaluate whether users notice the 'via' indicator, whether they understand the meaning of the 'via' indicator, and how their understanding affects what action users take with the email. We find that the majority of participants notice the 'via' indicator, but still proceed with unsafe behavior due to misunderstandings of what the indicator represents. Additionally, the understanding of what the 'via' indicator represents is heavily influenced by how familiar users are with the 'via' domain. The use of a more familiar and seemingly trustworthy domain thwarted the 'via' indicator's intended goal of conveying potential security concerns. These findings highlight the shortcomings of current passive security indicator design, and emphasize the need for indicators that users will notice and understand while still providing salient security information.

Acknowledgments

We thank our anonymous reviewers for their constructive feedback that helps make this paper better. We thank Kristen Vaccaro and Mary Anne Smart for their help with designing the study. We thank Cindy Moore and Jennifer Folkestad for their operational support. Funding for this work was provided in part by National Science Foundation grant CNS-2152644, the UCSD CSE Postdoctoral Fellows program, the Irwin Mark and Joan Klein Jacobs Chair in Information and Computer Science.

References

[1] Saeed Abu-Nimeh, Dario Nappa, Xinlei Wang, and Suku Nair. A Comparison of Machine Learning Techniques for Phishing Detection. In *Proceedings of the Anti-phishing Working Groups 2nd Annual Ecrime Researchers Summit*, pages 60–69, 2007.

[2] Devdatta Akhawe and Adrienne Porter Felt. Alice in Warningland: A Large-Scale Field Study of Browser Security Warning Effectiveness. In *Proceedings of the 22nd USENIX Security Symposium*, 2013.

[3] Sara Albakry, Kami Vaniea, and Maria K. Wolters. What Is This Url's Destination? Empirical Evaluation of Users' Url Reading. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2020.

[4] Kholoud Althobaiti, Nicole Meng, and Kami Vaniea. I Don't Need an Expert! Making Url Phishing Features Human Comprehensible. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2021.

[5] Chase Bank. Frequently Asked Questions: Fraud, 01 2023. https://www.chase.com/digital/resources/privacy-security/questions/fraud.

[6] Maxim Baryshevtsev and Joseph McGlynn. Persuasive Appeals Predict Credibility Judgments of Phishing Messages. *Cyberpsychology, Behavior, and Social Networking*, 23(5):297–302, 2020.

[7] Simon Bell and Peter Komisarczuk. An Analysis of Phishing Blacklists: Google Safe Browsing, Openphish, and Phishtank. In *Proceedings of the 2020 Australasian Computer Science Week Multiconference*, pages 1–11, 2020.

[8] Zinaida Benenson, Freya Gassmann, and Robert Landwirth. Unpacking Spear Phishing Susceptibility. In *Proceedings of the 2017 International Conference on Financial Cryptography and Data Security*, pages 610–627. Springer, 2017.

[9] Mark Blythe, Helen Petrie, and John A. Clark. F for Fake: Four Studies on How We Fall for Phish. In *Proceedings of the 2011 CHI Conference on Human Factors in Computing Systems*, pages 3469–3478, 2011.

[10] Paolo Buono, Giuseppe Desolda, Francesco Greco, and Antonio Piccinno. Let Warnings Interrupt the Interaction and Explain: Designing and Evaluating Phishing Email Warnings. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–6, 2023.

[11] Deanna D. Caputo, Shari Lawrence Pfleeger, Jesse D. Freeman, and M. Eric Johnson. Going Spear Phishing: Exploring Embedded Training and Awareness. *IEEE Security & Privacy*, 2013.

[12] Jianjun Chen, Vern Paxson, and Jian Jiang. Composition Kills: A Case Study of Email Sender Authentication. In *Proceedings of the 29th USENIX Security Symposium*, pages 2183–2199, 2020.

[13] Debra L. Cook, Vijay K. Gurbani, and Michael Daniluk. Phishwish: A Simple and Stateless Phishing Filter. *Security and Communication Networks*, 2009.

[14] Marco De Bona and Federica Paci. A Real World Study on Employees' Susceptibility to Phishing Attacks. In *Proceedings of the 15th International Conference on Availability, Reliability and Security*, pages 1–10, 2020.

[15] Rachna Dhamija and J. D. Tygar. The Battle Against Phishing: Dynamic Security Skins. In *Proceedings of the 2005 Symposium on Usable Privacy and Security*, pages 77–88, 2005.

[16] Rachna Dhamija, J. D. Tygar, and Marti Hearst. Why Phishing Works. In *Proceedings of the 2006 CHI Conference on Human Factors in Computing Systems*, pages 581–590, 2006.

[17] Alejandra Diaz, Alan T. Sherman, and Anupam Joshi. Phishing in an Academic Community: A Study of User Susceptibility and Behavior. *Cryptologia*, 2020.

[18] Julie S. Downs, Mandy Holbrook, and Lorrie Faith Cranor. Behavioral Response to Phishing Risk. In *Proceedings of the Anti-Phishing Working Groups 2nd Annual eCrime Researchers Summit*, pages 37–44, 2007.

[19] Julie S. Downs, Mandy B. Holbrook, and Lorrie Faith Cranor. Decision Strategies and Susceptibility to Phishing. In *Proceedings of the Second Symposium on Usable Privacy and Security*, pages 79–90, 2006.

[20] Sevtap Duman, Kubra Kalkan-Cakmakci, Manuel Egele, William Robertson, and Engin Kirda. Emailprofiler: Spearphishing Filtering With Header and Stylometric Features of Emails. In *Proceedings of the 2016 IEEE Annual Computer Software and Applications Conference*, pages 408–416, 2016.

[21] EasyDMARC. Email Forwarding and DMARC DKIM SPF, 05 2022. https://easydmarc.com/blog/email-forwarding-and-dmarc-dkim-spf/.

[22] Serge Egelman, Lorrie Faith Cranor, and Jason Hong. You've Been Warned: An Empirical Study of the Effectiveness of Web Browser Phishing Warnings. In *Proceedings of the 2009 CHI Conference on Human Factors in Computing Systems*, pages 1065–1074, 2008.

[23] Franz Faul, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. Statistical Power Analyses Using G* Power 3.1: Tests for Correlation and Regression Analyses. *Behavior research methods*, 41(4):1149–1160, 2009.

[24] Adrienne Porter Felt, Elizabeth Ha, Serge Egelman, Ariel Haney, Erika Chin, and David Wagner. Android Permissions: User Attention, Comprehension, and Behavior. In *Proceedings of the Eighth Symposium on Usable Privacy and Security*, pages 1–14, 2012.

[25] Ian Fette, Norman Sadeh, and Anthony Tomasic. Learning to Detect Phishing Emails. In *Proceedings of the 16th International Conference on World Wide Web*, pages 649–656, 2007.

[26] Anjuli Franz, Verena Zimmermann, Gregor Albrecht, Katrin Hartwig, Christian Reuter, Alexander Benlian, Joachim Vogt, et al. SOK: Still Plenty of Phish in the Sea-a Taxonomy of User-Oriented Phishing Interventions and Avenues for Future Research. In *Proceedings of Seventeenth Symposium on Usable Privacy and Security*, pages 339–358, 2021.

[27] Sujata Garera, Niels Provos, Monica Chew, and Aviel D. Rubin. A Framework for Detection and Measurement of Phishing Attacks. In *Proceedings of the 2007 ACM Workshop on Recurring Malcode*, pages 1–8, 2007.

[28] Google. Google Safe Browsing, 01 2023. https://safebrowsing.google.com/.

[29] Ayako Akiyama Hasegawa, Naomi Yamashita, Mitsuaki Akiyama, and Tatsuya Mori. Why They Ignore English Emails: The Challenges of Non-Native Speakers in Identifying Phishing Emails. In *Proceedings of Seventeenth Symposium on Usable Privacy and Security*, pages 319–338, 2021.

[30] Farkhondeh Hassandoust, Harminder Singh, and Jocelyn Williams. The Role of Contextualization in Individuals' Vulnerability to Phishing Attempts. *Australasian Journal of Information Systems*, 2020.

[31] Ryan Heartfield and George Loukas. A Taxonomy of Attacks and a Survey of Defence Mechanisms for Semantic Social Engineering Attacks. *ACM Computing Surveys*, 2015.

[32] Cormac Herley. So Long, and No Thanks for the Externalities: The Rational Rejection of Security Advice by Users. In *Proceedings of the 2009 Workshop on New Security Paradigms Workshop*, pages 133–144, 2009.

[33] Amir Herzberg and Ahmad Gbara. Trustbar: Protecting (Even Naive) Web Users From Spoofing and Phishing Attacks. Technical report, 2004. http://eprint.iacr.org/2004/155.

[34] Grant Ho, Aashish Sharma, Mobin Javed, Vern Paxson, and David Wagner. Detecting Credential Spearphishing in Enterprise Settings. In *Proceedings of the 2017 USENIX Security Symposium*, pages 469–485, 2017.

[35] Hang Hu and Gang Wang. End-to-End Measurements of Email Spoofing Attacks. In *Proceedings of the 27th USENIX Security Symposium*, pages 1095–1112, 2018.

[36] Collin Jackson, Daniel R. Simon, Desney S. Tan, and Adam Barth. An Evaluation of Extended Validation and Picture-in-Picture Phishing Attacks. In *Proceedings of the 2007 International Conference on Financial Cryptography and Data Security*, pages 281–293. Springer, 2007.

[37] Daniel Jampen, Gürkan Gür, Thomas Sutter, and Bernhard Tellenbach. Don't Click: Towards an Effective Anti-Phishing Training. A Comparative Literature Review. *Human-centric Computing and Information Sciences*, 2020.

[38] Yogesh Joshi, Samir Saklikar, Debabrata Das, and Subir Saha. Phishguard: A Browser Plug-in for Protection From Phishing. In *Proceedings of the 2008 International Conference on Internet Multimedia Services Architecture and Applications*, pages 1–6. IEEE, 2008.

[39] Smirity Kaushik, Yaxing Yao, Pierre Dewitte, and Yang Wang. "How I Know for Sure": People's Perspectives on Solely Automated Decision-Making (SADM). In *Proceedings of the Seventeenth Symposium on Usable Privacy and Security*, pages 159–180, 2021.

[40] Mahmoud Khonji, Youssef Iraqi, and Andrew Jones. Mitigation of Spear Phishing Attacks: A Content-Based Authorship Identification Framework. In *Proceedings of the 2011 International Conference for Internet Technology and Secured Transactions*, 2011.

[41] Ponnurangam Kumaraguru, Justin Cranshaw, Alessandro Acquisti, Lorrie Cranor, Jason Hong, Mary Ann Blair, and Theodore Pham. School of Phish: A Real-World Evaluation of Anti-Phishing Training. In *Proceedings of the Fifth Symposium on Usable Privacy and Security*, pages 1–12, 2009.

[42] Ponnurangam Kumaraguru, Yong Rhee, Alessandro Acquisti, Lorrie Faith Cranor, Jason Hong, and Elizabeth Nunge. Protecting People From Phishing: The Design and Evaluation of an Embedded Training Email System. In *Proceedings of the 2007 CHI Conference on Human Factors in Computing Systems*, pages 905–914, 2007.

[43] Ponnurangam Kumaraguru, Yong Rhee, Steve Sheng, Sharique Hasan, Alessandro Acquisti, Lorrie Faith Cranor, and Jason Hong. Getting Users to Pay Attention to Anti-Phishing Education: Evaluation of Retention and Transfer. In *Proceedings of the Anti-phishing Working Groups 2nd Annual eCrime Researchers Summit*, 2007.

[44] Ponnurangam Kumaraguru, Steve Sheng, Alessandro Acquisti, Lorrie Faith Cranor, and Jason Hong. Lessons From a Real World Evaluation of Anti-Phishing Training. In *Proceedings of the Anti-phishing Working Groups 3rd Annual eCrime Researchers Summit*, 2008.

[45] Ponnurangam Kumaraguru, Steve Sheng, Alessandro Acquisti, Lorrie Faith Cranor, and Jason Hong. Teaching Johnny Not to Fall for Phish. *ACM Transactions on Internet Technology*, 2010.

[46] Elmer Lastdrager, Inés Carvajal Gallardo, Pieter Hartel, and Marianne Junger. How Effective Is Anti-Phishing Training for Children? In *Proceedings of the Thirteenth Symposium on Usable Privacy and Security*, pages 229–239, 2017.

[47] Kenneth R. Laughery and Michael S. Wogalter. Designing Effective Warnings. *Reviews of Human Factors and Ergonomics*, 2006.

[48] Sarah Lichtenstein, Baruch Fischhoff, and Lawrence D. Phillips. Calibration of Probabilities: The State of the Art. In *Proceedings of the Fifth Research Conference on Subjective Probability, Utility, and Decision Making*, pages 275–324, 1977.

[49] Eric Lin, Saul Greenberg, Eileah Trotter, David Ma, and John Aycock. Does Domain Highlighting Help People Identify Phishing Sites? In *Proceedings of the 2011 CHI Conference on Human Factors in Computing Systems*, pages 2075–2084, 2011.

[50] Tian Lin, Daniel E. Capecci, Donovan M. Ellis, Harold A. Rocha, Sandeep Dommaraju, Daniela S. Oliveira, and Natalie C. Ebner. Susceptibility to Spear-Phishing Emails: Effects of Internet User Demographics and Email Content. *ACM Transactions on Computer-Human Interaction*, 2019.

[51] Enze Liu, Gautam Akiwate, Mattijs Jonker, Ariana Mirian, Grant Ho, Geoffrey M. Voelker, and Stefan Savage. Forward Pass: On the Security Implications of Email Forwarding Mechanism and Policy. In *Proceedings of the 8th IEEE European Symposium on Security and Privacy*, 2023.

[52] Enze Liu, Gautam Akiwate, Mattijs Jonker, Ariana Mirian, Stefan Savage, and Geoffrey M. Voelker. Who's Got Your Mail? Characterizing Mail Service Provider Usage. In *Proceedings of the 21st ACM Internet Measurement Conference*, pages 122–136, 2021.

[53] Gang Liu, Guang Xiang, Bryan A. Pendleton, Jason I. Hong, and Wenyin Liu. Smartening the Crowds: Computational Techniques for Improving Human Verification to Fight Phishing Scams. In *Proceedings of the Seventh Symposium on Usable Privacy and Security*, pages 1–13, 2011.

[54] Samuel Marchal, Giovanni Armano, Tommi Gröndahl, Kalle Saari, Nidhi Singh, and N Asokan. Off-the-Hook: An Efficient and Usable Client-Side Phishing Prevention Application. *IEEE Transactions on Computers*, 2017.

[55] Samuel Marchal, Kalle Saari, Nidhi Singh, and N Asokan. Know Your Phish: Novel Techniques for Detecting Phishing Sites and Their Targets. In *Proceedings of the 36th International Conference on Distributed Computing Systems*, pages 323–333, 2016.

[56] Mary L. McHugh. Interrater Reliability: The Kappa Statistic. *Biochemia medica*, 2012.

[57] Gregory D. Moody, Dennis F. Galletta, and Brian Kimball Dunn. Which Phish Get Caught? An Exploratory Study of Individuals' Susceptibility to Phishing. *European Journal of Information Systems*, 2017.

[58] James Nicholson, Lynne Coventry, and Pam Briggs. Can We Fight Social Engineering Attacks by Social Means? Assessing Social Salience as a Means to Improve Phish Detection. In *Proceedings of the Thirteenth Symposium on Usable Privacy and Security*, pages 285–298, 2017.

[59] James Nicholson, Yousra Javed, Matt Dixon, Lynne Coventry, Opeyemi Dele Ajayi, and Philip Anderson. Investigating Teenagers' Ability to Detect Phishing Messages. In *Proceedings of the 2020 IEEE European Symposium on Security and Privacy Workshops*, 2020.

[60] Daniela Oliveira, Harold Rocha, Huizi Yang, Donovan Ellis, Sandeep Dommaraju, Melis Muradoglu, Devon Weir, Adam Soliman, Tian Lin, and Natalie Ebner. Dissecting Spear Phishing Emails for Older vs Young Adults: On the Interplay of Weapons of Influence and Life Domains in Predicting Susceptibility to Phishing. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 6412–6424, 2017.

[61] Justin Petelka, Yixin Zou, and Florian Schaub. Put Your Warning Where Your Link Is: Improving and Evaluating Email Phishing Warnings. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2019.

[62] PhishTank. Phishtank, 01 2023. https://phishtank.org/.

[63] ProofPoint. Phishing Protection, 01 2023. https://www.proofpoint.com/us/solutions/protect-against-phishing.

[64] Benjamin Reinheimer, Lukas Aldag, Peter Mayer, Mattia Mossano, Reyhan Duezguen, Bettina Lofthouse, Tatiana Von Landesberger, and Melanie Volkamer. An Investigation of Phishing Awareness and Education Over Time: When and How to Best Remind Users. In *Proceedings of the Sixteenth Symposium on Usable Privacy and Security*, pages 259–284, 2020.

[65] Markus Riek, Rainer Bohme, and Tyler Moore. Measuring the Influence of Perceived Cybercrime Risk on Online Service Avoidance. *IEEE Transactions on Dependable and Secure Computing*, 2015.

[66] Stefan A. Robila and James W. Ragucci. Don't Be a Phish: Steps in User Education. *ACM SIGCSE Bulletin*, 2006.

[67] Martina Angela Sasse, Sacha Brostoff, and Dirk Weirich. Transforming the 'Weakest Link'—a Human/Computer Interaction Approach to Usable and Effective Security. *BT Technology Journal*, 2001.

[68] Katharina Schiller, Florian Adamsky, and Zinaida Benenson. Towards an Empirical Study to Determine the Effectiveness of Support Systems Against E-Mail Phishing Attacks. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2023.

[69] Skipper Seabold and Josef Perktold. Statsmodels: Econometric and Statistical Modeling With Python. In *Proceedings of the 9th Python in Science Conference*, 2010.

[70] Kaiwen Shen, Chuhan Wang, Minglei Guo, Xiaofeng Zheng, Chaoyi Lu, Baojun Liu, Yuxuan Zhao, Shuang Hao, Haixin Duan, Qingfeng Pan, et al. Weak Links in Authentication Chains: A Large-Scale Analysis of Email Sender Spoofing Attacks. In *Proceedings of the 2021 USENIX Security Symposium*, 2021.

[71] Steve Sheng, Mandy Holbrook, Ponnurangam Kumaraguru, Lorrie Faith Cranor, and Julie Downs. Who Falls for Phish? A Demographic Analysis of Phishing Susceptibility and Effectiveness of Interventions. In *Proceedings of the 2010 CHI conference on Human Factors in Computing Systems*, pages 373–382, 2010.

[72] Steve Sheng, Bryant Magnien, Ponnurangam Kumaraguru, Alessandro Acquisti, Lorrie Faith Cranor, Jason Hong, and Elizabeth Nunge. Anti-Phishing Phil: The Design and Evaluation of a Game That Teaches People Not to Fall for Phish. In *Proceedings of the Third Symposium on Usable Privacy and Security*, pages 88–99, 2007.

[73] Kuldeep Singh, Palvi Aggarwal, Prashanth Rajivan, and Cleotilde Gonzalez. Training to Detect Phishing Emails: Effects of the Frequency of Experienced Phishing Emails. In *Proceedings of the 2019 Human Factors and Ergonomics Society Annual Meeting*, 2019.

[74] Eric Spero and Robert Biddle. Out of Sight, Out of Mind: Ui Design and the Inhibition of Mental Models of Security. In *Proceedings of the 2020 New Security Paradigms Workshop*, pages 127–143, 2020.

[75] Gianluca Stringhini and Olivier Thonnard. That Ain't You: Blocking Spearphishing Through Behavioral Modelling. In *Proceedings of the 2015 International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment*, pages 78–97, 2015.

[76] Jenny Tang, Eleanor Birrell, and Ada Lerner. Replication: How Well Do My Results Generalize Now? The External Validity of Online Privacy and Security Surveys. In *Proceedings of the Eighteenth Symposium on Usable Privacy and Security*, pages 367–385, 2022.

[77] NCRLY Teraguchi and John C. Mitchell. Client-Side Defense Against Web-Based Identity Theft. *Computer Science Department, Stanford University*, 2004.

[78] Arun Vishwanath, Brynne Harrison, and Yu Jie Ng. Suspicion, Cognition, and Automaticity Model of Phishing Susceptibility. *Communication Research*, pages 1146–1166, 2018.

[79] Arun Vishwanath, Tejaswini Herath, Rui Chen, Jingguo Wang, and H. Raghav Rao. Why Do People Get Phished? Testing Individual Differences in Phishing Vulnerability Within an Integrated, Information Processing Model. *Decision Support Systems*, 2011.

[80] Melanie Volkamer, Karen Renaud, and Paul Gerber. Spot the Phish by Checking the Pruned URL. *Information & Computer Security*, 2016.

[81] Melanie Volkamer, Karen Renaud, Benjamin Reinheimer, and Alexandra Kunz. User Experiences of Torpedo: Tooltip-Powered Phishing Email Detection. *Computers & Security*, 2017.

[82] Chenkai Wang and Gang Wang. Revisiting Email Forwarding Security Under the Authenticated Received Chain Protocol. In *Proceedings of the 2022 ACM Web Conference*, pages 681–689, 2022.

[83] Ge Wang, He Liu, Sebastian Becerra, Kai Wang, Serge J Belongie, Hovav Shacham, and Stefan Savage. Verilogo: Proactive Phishing Detection via Logo Recognition. *Department of Computer Science and Engineering, University of California*, 2011.

[84] Rick Wash. How Experts Detect Phishing Scam Emails. *Proceedings of the ACM on Human-Computer Interaction*, 2020.

[85] Rick Wash and Molly M Cooper. Who Provides Phishing Training? Facts, Stories, and People Like Me. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2018.

[86] Rick Wash, Norbert Nthala, and Emilee Rader. Knowledge and Capabilities That Non-Expert Users Bring to Phishing Detection. In *Proceedings of Seventeenth Symposium on Usable Privacy and Security*, 2021.

[87] Zikai Alex Wen, Zhiqiu Lin, Rowena Chen, and Erik Andersen. What. Hack: Engaging Anti-Phishing Training Through a Role-Playing Phishing Simulation Game. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019.

[88] Liu Wenyin, Ning Fang, Xiaojun Quan, Bite Qiu, and Gang Liu. Discovering Phishing Target Based on Semantic Link Network. *Future Generation Computer Systems*, 2010.

[89] Colin Whittaker, Brian Ryner, and Marria Nazif. Large-Scale Automatic Classification of Phishing Pages. In *Proceedings of the 2010 Network and Distributed System Security Symposium*, 2010.

[90] Wikipedia. Kruskal-Wallis OneWay Analysis of Variance, 06 2023. https://en.wikipedia.org/wiki/Kruskal%E2%80%93Wallis_one-way_analysis_of_variance.

[91] Emma J Williams and Danielle Polage. How Persuasive Is Phishing Email? The Role of Authentic Design, Influence and Current Events in Email Judgements. *Behaviour and Information Technology*, 2019.

[92] Ryan T. Wright and Kent Marett. The Influence of Experiential and Dispositional Factors in Phishing: An Empirical Investigation of the Deceived. *Journal of Management Information Systems*, 2010.

[93] Min Wu, Robert C. Miller, and Simson L. Garfinkel. Do Security Toolbars Actually Prevent Phishing Attacks? In *Proceedings of the 2006 CHI Conference on Human Factors in Computing Systems*, 2006.

[94] Guang Xiang, Jason Hong, Carolyn P. Rose, and Lorrie Cranor. Cantina+ a Feature-Rich Machine Learning Framework for Detecting Phishing Web Sites. *ACM Transactions on Information and System Security*, 2011.

[95] Weining Yang, Aiping Xiong, Jing Chen, Robert W. Proctor, and Ninghui Li. Use of Phishing Training to Improve Security Warning Compliance: Evidence From a Field Experiment. In *Proceedings of the 2017 Hot Topics in Science of Security: Symposium and Bootcamp*, pages 52–61, 2017.

[96] Ka-Ping Yee and Kragen Sitaker. Passpet: Convenient Password Management and Phishing Protection. In *Proceedings of the Second Symposium on Usable Privacy and Security*, pages 32–43, 2006.

[97] Yaman Yu, Saidivya Ashok, Smirity Kaushi, Yang Wang, and Gang Wang. Design and Evaluation of Inclusive Email Security Indicators for People With Visual Impairments. In *Proceedings of the 2023 IEEE Symposium on Security and Privacy*, 2023.

[98] Yue Zhang, Serge Egelman, Lorrie Cranor, and Jason Hong. Phinding Phish: Evaluating Anti-Phishing Tools. *Carnegie Mellon University*, 2007.

[99] Yue Zhang, Jason I. Hong, and Lorrie F. Cranor. Cantina: A Content-Based Approach to Detecting Phishing Web Sites. In *Proceedings of the 16th International Conference on World Wide Web*, 2007.

[100] Sarah Zheng and Ingolf Becker. Presenting Suspicious Details in User-Facing E-Mail Headers Does Not Improve Phishing Detection. In *Proceedings of the Eighteenth Symposium on Usable Privacy and Security*, pages 253–271, 2022.

A Appendix

A.1 Survey

Prescreening Questions

1. Are you currently a Gmail user? [Yes/No]
2. How long have you been using Gmail? [Less than a year - Four or more years]

Survey Questions

1. Please provide the name of the person or entity that sent the email?
2. How confident are you that your answer is correct? [1-5]
3. Please provide the email address of the person or entity that sent the email?
4. How confident are you that your answer is correct? [1-5]
5. How did you decide who the sender of the email was?
6. Do you remember seeing something similar to what is highlighted in the picture shown below during the study? (yes & I know what it means/I don't know what it means or no)
7. Please elaborate on what you think 'via' means
8. What information do you think Gmail is trying to communicate by showing 'via' for this email? Please make your best guess and feel free to refer back to the email.
9. Why do you think Gmail has chosen to display 'via' to users for certain emails? Please make your best guess and feel free to refer back to the email.
10. Chase.com used or instructed chasesupport.com to send the email [agreement]
11. Please explain your choice

A.2 Codebook for Question

Table 5 and Table 6 show themes, codes and explanations for the questions that asking for interpretation of 'via'. This codebook coded the combined answer for the question "Please elaborate on what you think 'via' means" and " What information do you think Gmail is trying to communicate by showing 'via' for this email? "

Theme	Code	Explanation
Via indicates sender	Identifying the actual sender	Indicating or Identifying the actual sender (i.e., this is coming from chasesupport.com that's it)
Via indicates group association	From Chase's support division	Coming from chase' support division or mentioning the relationship between chase.com and chasesupport.com)
Via means through	Through a third party	Identifying that the email is sent through a third party or mailing list

Table 5: Codebook on questions asking for participants' interpretation of 'via' in the Support group

Table 7 and Table 8 below show themes, codes and explanations for the question "Why do you think Gmail has chosen to display 'via' to users for certain emails?" on the Support group and the Random group.

Theme	Code	Explanation
Via indicates sender	Identifying the actual sender	Indicating or identifying the actual sender or server that sent the emails
Via indicates group association	From Chase's support division	Coming from chase' support division (i.e. mentioning the relationship between chase.com and chasesupport.com)
Via encourages caution	Security or scam	Showing via to notify security reasons
	Forward address	Showing the forward address
Via means through	Through a third party	Showing that the email is sending through a third party or a mailing list
	Return-path domain	Identifying the domain as a return-path domain
Others	From a new contact	Identifying the email is from a new contact

Table 6: Codebook on questions asking for participants' interpretation of 'via' in the Random group

Theme	Code	Explanation
Gmail displays via for safety	Security	Showing for security reasons like phishing and legitimacy
Gmail displays via to inform you about the email	Outsourced	Specifically mentioning that the email is sent by a third party and not directly from Chase
	More info	Providing more information (e.g. where it comes from or on the sender) or showing to provide more transparency
Gmail always displays via	Showing reasons	Showing users why they are getting the email.
	Explaining via	Explaining that the email is sent on behalf of other sender
Unsure why Gmail displays via	Unsure	Don't know or unsure

Table 7: Codebook for the question: "Why do you think Gmail has chosen to display via" for the Support group

Theme	Code	Explanation
Gmail displays via for safety	Security	Showing for security reasons like phishing and legitimacy
Gmail displays via to inform you about the email	Outsourced	Specifically mentioning that the email is sent by a third party and not directly from Chase
	More info	Providing more information (e.g. where it comes from or on the sender) or showing to provide more transparency
	Authenticity	Adding the domain to show authenticity
Gmail always displays via	Explaining via	Explaining that the email is sent on behalf of other sender
Unsure why Gmail displays via	Unsure	Don't know or not sure

Table 8: Codebook for the question: Why do you think Gmail has chosen to display via for the Random group

Table 9 and Table 10 below show themes, codes and explanations for the reasons on the agreement questions "Chase.com used or instructed chasesupport.com to send the email" on the Support group and the Random group.

Theme	Code	Explanation
The via domain is suspicious	Scam	Email could be a scam or phishing email
	Suspicious	Email looks suspicious or the user feels concerned
That's just how it works	Sending service	The user identifies the random domain as a third party sending service or a bot
	Relevant domain	The domain name (chasesupport.com) seems related to chase
	Verified by Google	Google or Gmail verified that this email
	Via meaning	It conforms to the definition of 'via'
	Chase initiated	Chase initiated or approved the email correspondence
	Make sense	The explanation makes sense or conforms to users' mental model
	Unable to tell	The user cannot determine the relationship between chase and chasesupport. The user only knows that the actual sender is different than the purported sender
I'm not sure who did what	Unsure	Don't know or not sure

Table 9: Codebook for the reasons that participants agree or not agree that chase instructed the entity to send the email for the Support group

Theme	Code	Explanation
The via domain is suspicious	Scam	Email could be a scam or phishing email
	Suspicious	Email looks suspicious or the user feels concerned
That's just how it works	Sending service	The user identifies the random domain as a third party sending service or a bot
	The meaning of via	It conforms to the definition of 'via'
	Chase initiated	Chase initiated or approved the email correspondence
	Make sense	The explanation makes sense or conforms to users' mental model
I'm not sure who did what	Unsure	Don't know or unsure
Chase doesn't do this	Uncommon	It's an uncommon case or domain
	The other way around	The other way around (random instructed chase to send the email)

Table 10: Codebook for the reasons that participants agree or not agree that chase instructed the entity to send the email for the Random group

‘Give Me Structure’: Synthesis and Evaluation of a (Network) Threat Analysis Process Supporting Tier 1 Investigations in a Security Operation Center

Leon Kersten
Eindhoven University of Technology

Tom Mulders
Eindhoven University of Technology

Emmanuele Zambon
Eindhoven University of Technology

Chris Snijders
Eindhoven University of Technology

Luca Allodi
Eindhoven University of Technology

Abstract

Current threat analysis processes followed by tier-1 (T1) analysts in a Security Operation Center (SOC) rely mainly on tacit knowledge, and can differ greatly across analysts. The lack of structure and clear objectives to T1 analyses makes operative inefficiencies hard to spot, SOC performance hard to measure (and therefore improve), results in overall lower security for the monitored environment(s), and contributes to analyst burnout. In this work we collaborate with a commercial SOC to devise a 4-stage (network) threat analysis process to support the collection and analysis of relevant information for threat analysis. We conduct an experiment with ten T1 analysts employed in the SOC and show that analysts following the proposed process are 2.5 times more likely to produce an accurate assessment than analysts who do not. We evaluate qualitatively the effects of the process on analysts decisions, and discuss implications for practice and research.

1 Introduction

As the volume and sophistication of cyber-attacks increase, the security of networks and systems is of key societal and economic importance. *Security Operation Centers* (SOCs) are business units (within a larger organizational setting), or services (that typically sell managed security services such as security monitoring to third party organizations) whose purpose is to detect cyber-attacks within the monitored environments. Their effectiveness is of primary importance both operationally (to maintain security) and strategically (to deter attacks) [6, 31]. A typical SOC is structured around a tiered system of analysts whereby incoming security events in the form of alerts are first analyzed by tier 1 (T1) analysts, who are typically junior and relatively inexperienced [20] and only escalated to higher tiers (typically through T2 and up to T3 analysts) when the T1 believes the event to be a potential threat to an organization [14]. This tiered system generates a procedure whereby T1 analysts analyze plentiful of false positive alerts (i.e., that are ‘not interesting’ for escalation) [2] and pass on relevant information to higher tiers for more in-depth investigation on alerts for which T1s cannot rule out evidence of attack [14].

Thus, the timeliness, accuracy, and relevancy of the information T1 analysts pass on to T2/T3 analysts is crucial to effective and efficient SOC operations, and by extension to the security of the monitored environments. Despite this, in many SOCs, the actions that SOC analysts take and the information they seek to inform their decisions depends mainly on tacit knowledge and their own background [5, 23], as opposed to a clear structure or framework to identify relevant evidence leading to well-informed decisions. T1 analysts typically do receive training, but generally in the form of internal procedures, systems, and new vulnerabilities [22], rather than in the form of an evidence-based decision-making process for effective threat analysis. This can lead to large differences in accuracy across T1 analysts [26, 28]. An unstructured analysis process can also be problematic in terms of the information passed over to a T2 analyst when an alert is escalated. T2 analysts then need to process reports that are less standardized, coherent and actionable.

Partially mitigating this issue, most SOCs implement so-called ‘playbooks’ or ‘runbooks’, documenting the procedures T1 analysts should follow when analyzing alerts related to a certain use case (e.g. a use case for ‘ransomware’). However, playbook documents are known to have update and maintenance issues, or only cover generic use cases that may induce analysts to use them less than originally intended [5]. In other extreme cases, some managed SOCs employ playbooks that are, in essence, automation rules to report information back to their customers. In those SOCs the T1 analyst (if present

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.
USENIX Symposium on Usable Privacy and Security (SOUPS) 2023.
August 6–8, 2023, Anaheim, CA, USA

at all) is essentially reduced to an automaton executing a specific algorithm for each type of alert [22]. This is problematic considering the dynamic nature of cyber-attacks, and the high rate of false positives generated by detectors [2]. Indeed, the relevant literature suggests that analysts perform better when trained on a large variety of threats whose investigation require high cognitive engagement, as opposed to executing pre-determined tasks [8]. In addition, automaton execution possibly leads to a higher likelihood of burnout [22]. For these combined reasons, numerous previous studies have stressed the importance of humans and their decision making abilities in a SOC [10, 22, 31]. In this work we focus on supporting T1 analysts' cognitive engagement by proposing a general framework ('structure') to guide the T1 through the threat analysis process. More specifically, we collaborate with a commercial SOC (the Eindhoven Security Hub SOC [1]) to devise, together with three senior analysts, a structure for the T1 threat analysis process. We then evaluate the effects of the proposed process via a controlled experiment with 10 T1 analysts recruited at the SOC jointly analyzing 200 alerts from a real monitored environment.

Scope and contribution. The aim of this paper is to evaluate whether the creative cognitive process behind threat analysis can be aided by providing analysts with guidance on what information to collect to address specific 'stages' of the threat analysis process. This is different from building an 'algorithm' or set of heuristics to automate analyst decisions, which is not within the scope of this paper. In terms of scope of the proposed process, we focus on network event analysis.

The remainder of the paper is structured as follows. Section 2 introduces the background on the role of T1 analysts in SOCs, their security analysis process, and discusses related work. Section 3 presents the research questions addressed in this work, and Section 4 describes the methodology to answer them. Section 5 presents our baseline threat analysis process in detail and Section 6 provides the results to validate our process. Finally, Section 7 discusses our findings and Section 8 provides conclusions.

2 Background and Related Work

2.1 SOCs and tier 1 analysts

A security operation center (SOC) is a provider of security services to organizations. SOCs can be internal, where they provide security services to their own (often large) organization, or external by providing security services to third parties. The security services provided by SOCs can range from network intrusion detection systems (NIDS), to endpoint detection and firewall monitoring. The tools providing these services may utilize static detection mechanisms, dynamic systems, machine learning, artificial intelligence and more. However, most of these technical security solutions in an operational SOC generate security events which are in the first instance evaluated by a T1 analyst in the SOC. The analysis performed by T1 analysts aims to discern interesting security events from not interesting events. In other words, identify 'interesting' security events to escalate to T2 and T3 security analysts for further investigation, communication to customers and possible mitigation actions.

Since T1 analysts are the first to analyze and classify security events, the accuracy and timeliness of their analysis is fundamental to a SOC. Indeed, an accurate and timely identification of a cyber-attack minimizes the time available for the attack to complete [8], or may prevent its impact to be fully realized. Unfortunately, investigations performed by T1 analysts are known to be, in general, error prone and time-consuming, despite being repetitive [10, 12, 22, 31]. Naturally, the quality of the analysis and thus the accuracy of the classification is dependent on the individual skill of the analyst. However, external factors can impact this significantly, such as the arrival of external vulnerability information, or the addition of a new network segment in scope of an analyst's monitoring. Given that this occurs regularly, yet not necessarily predictably, the quality of security event analyses can be quite variable [26].

2.2 The security analysis process

The analysis process of a T1 analyst has as input a security event, and as output a classification of this security event. Regardless of internal taxonomies, analysts can in general assign a security event to one of two groups: alerts worthy of escalation to a higher tier for further investigation (further referred to as 'interesting' alerts), and those who should not be escalated (further referred to as 'not interesting' alerts) [10].[2] To arrive at this conclusion, the analyst's job is to look for evidence relating to the security event under analysis with other (security or network) events from any available and relevant source, with the goal of performing a triage for a possible escalation to higher tiers [6, 19, 29].

The T1 analysts' workflow takes as input a large volume of information that a SOC analyst has to account for in order to classify a security alert. They may look at the source and destination IP addresses [7, 23], at an increase in activities on a certain network port [7], or at the packet size and the content of a payload [7]. From the literature we identify four main categories of information that is considered by an analyst: *Relevance indicators*, evaluating whether an alert is relevant to the scope of the analysis [6]; *Additional alerts*, considering whether other evidence exists that an attack may be ongoing [6, 7]; *Contextual information*, evaluating whether some evidence of an attack is present at or around the affected hosts or systems [7, 21]; *Attack Evidence*, evaluating whether

[1] https://www.eindhovensecurityhub.nl/

[2] More fine-grained evaluations (e.g., Command & Control traffic, suspicious/benign scanning activity, ..) are always possible and commonly employed in SOCs as 'metadata' attached to the categorization above.

there is evidence that the events generating the alert led to a (successful) attack [6]. Table 1 provides a summary.

On the other hand, different analysts are known to employ different strategies to analyse a specific alert [26, 28]. Indeed, there is no clear framework of reference on what information to collect and what evidence is relevant to which phase of the investigation [2]. This suggests that there is no one clear predefined process for T1 analysts to follow, and that analysis results are entirely left to an analyst's own background, knowledge, and skills [23]. This may lead to increased analyst burnout [14], and is particularly undesirable given the high-turnover nature of T1 analysts within SOCs (that are regularly substituted by more junior and inexperienced analysts).

2.3 Related Work

Most of the previous research on SOC analysts focuses on the work process of the SOC as a whole rather than specific roles within the SOC. Of the papers mentioned in this section, three are qualitative studies [6, 14, 23] and two include some quantitative results [28, 30]. Furthermore, previous work can be divided in those that specifically consider the alert analysis process of SOC analysts [6, 23, 28] (although with the abstraction level of SOCs as a whole) or those who consider the work of the analysts from an organizational perspective [14, 22].

D'Amico and Whitley [6] conducted a cognitive task analysis (CTA) on the general workflow of a SOC, the security analysis process and the decision making process of an analyst. The authors identified a tiered system where a large volume of data enters the SOC, and that in each tier data is either discarded or retained to transfer to the next tier. Their work identifies several pieces of information that SOC analysts utilize to analyze security events and based on the CTA the authors draw conclusions on how visualization tools can and should integrate in the SOC enviroment. Although their work does not conduct a CTA for specific roles within the SOC (e.g T1 analyst), it provides an overview of how the SOC as a whole analyze and handle incoming security events. Similarly, Zhong et al. [28], captured analysis operations performed by analysts in a SOC and the hypotheses they generate and utilize in this process. The authors conduct CTA to capture fine-grained processes performed as part of the alert analysis. Interestingly, the authors highlight the observation that analysts employ different strategies and processes to explore the data and generate hypotheses to investigate. In later work Zhong et al. [30] propose a tool that automates the data triage aspect of aT1 analyst's work. They observed high-performance and satisfactory false-positive rates. They do note, however, that the quality of the system depends on the quality of the triage traces, which in turn depends on the quality of the analyst. Notably, this approach utilizes the operations of the analysts, such as "searching","selecting" and "filtering" [30], and does not capture why an analyst performs this action, nor what evidence is obtained from this operation.

Kokulu et al. conducted a qualitative study on issues within the SOC [14]. One of the primary findings is the current metrics for SOC performance are not effective. Moreover, this is a point of contention between security analysts and their managers. They also found the speed of response and the level of automation to be similarly (and very) important for effective SOC operations. Additionally, they noted that poor analyst training and high false-positive rates are issues within the SOCs in their research. This all culminates into poor quality of analyses, if left unaddressed.

Another interesting observation noted by Sunderamurthy et al [23] is the problem of tacit knowledge within the SOC; decisions made by security analysts are based on intuition and not documented. Often, the security analysts cannot clearly communicate their knowledge related to the incident and the reason for their classification of this incident [23]. This is a key component of the services provided by SOCs, as the contact point for the monitored environment must be provided with accurate and actionable evidence of a security incident. The contact person must be convinced that mitigation is necessary and warrant a potential interruption of business processes. T1 security analysts must do this as well when escalating to T2 or T3 analysts. Good communication about what a T1 analyst has observed, supporting evidence and their decision process is therefore a must in an effective SOC. On this line, [23] reports that "*SOC jobs such as incident response and forensic analysis have become so sophisticated and expertise driven that understanding the process is nearly impossible without doing the job.*"

3 Problem Statement and Research Questions

The process of alert investigation is repetitive, time-consuming and error prone [10, 12, 22, 31]. Much research has been done on the automation of individual steps or parts of the investigation, such as correlation and alert reduction, often relying on automated learning techniques [24, 31]. Past research also provided an high level overview of the workflow of an analyst [6, 7, 28]. However, to our knowledge a clear structure of the investigative process, and an evaluation of the extent to which it would aid in accurate decision making by T1 analysts, is currently missing [25]. The problem statement above gives rise to the following two research questions:

RQ1: Which sequence of tasks and information gathering should a tier 1 analyst perform when executing a threat analysis process to analyze network security alerts?

RQ2: To what extent can the derived threat analysis process increase the accuracy of classifying network security alerts for tier 1 analysts?

Table 1: Categories of information SOC analysts employ to classify alerts

Information Category	Definition	References
Relevance indicators	Information to classify whether the alert under investigation is even relevant for the SOC, based on the signature and the scope of the customer.	[2, 6]
Additional alerts	Alerts related to the current alert that the analyst is investigating. This may be previous instances of the same alert triggering or alerts that surround the current alert.	[6, 7]
Contextual information	Information about the behavior and other observables of the involved internal host.	[2, 7, 12, 21]
Attack evidence	Any evidence relating to the alleged attack including the type of attack, attacker and any indication of success.	[2, 6]

4 Methodology

Overview of method. To answer our research questions we rely on an ongoing collaboration with a commercial (managed) SOC, the Eindhoven Security Hub SOC (for brevity referred to as 'the SOC'), providing network monitoring services for small and medium-size organizations active in education, IT-services, and manufacturing.

RQ1. To derive the threat analysis process we worked closely with a T2 security analysis expert with 4+ years of experience who is currently active in the SOC to identify which information a T1 analyst should consider for the 'escalation' of the alert to be useful for the higher-tier analysts. The derived information was then mapped to the categories presented in Table 1 and used to build a step-wise process for the analysis. This process was then iteratively and independently evaluated by two senior analysts (with respectively 15+ and 10+ years of experience) active in the SOC, until all three experts were in agreement on the resulting threat analysis process. Implementation details and results are given in Section 5.

RQ2. To validate the identified threat analysis process we designed an experiment to compare the performance of SOC analysts who employ the process to conduct analysis of alert data in the SOC, against that of analysts who do not. We employ sensor data from one of the organizations monitored by the SOC to sample alerts from a real-life environment. This ensures that baseline information (such as the IP space of that organization) is already known to the analysts. In addition to using alerts from our sensor, we generated additional alerts by injecting attacks into the virtual SOC environment to validate our process on alerts relating to successful attacks. To not affect SOC operations, we reproduce a near identical virtual environment that T1 analysts employ in the SOC in their day to day work (details of the environment can be found in Appendix B), and recruit our subjects from the pool of analysts employed at the time at the SOC.

4.1 Synthesis of the threat analysis process

Figure 1 provides an overview of the iterative process we employ to construct our proposed threat analysis process.

Preliminary alert analysis. In order to establish a set of information a T1 analyst should collect we adopted a bottom-up approach and sampled a set of 10 security alerts from a prototype sensor of the SOC. This relatively low number of alerts was chosen in first instance under the observation (and in consultation with the involved SOC experts) that the T1 analysis process is very repetitive and does not vary significantly across (network) security alerts. To make sure saturation in the collected information steps is reached, we adopted a step-wise process whereby additional information of relevance to the process is added to the set as each new alert is analyzed.

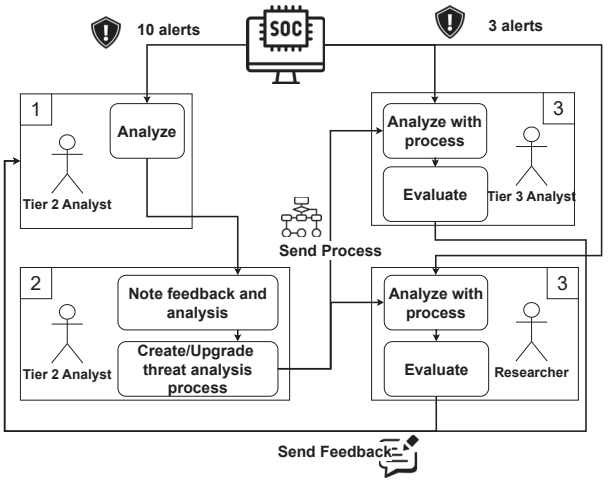


Figure 1: Synthesis of the baseline threat analysis process

The ten sampled alerts consist of malware, exploits, command & control, policy violations and scan alerts. The alerts were randomly selected from one of the monitored environments in the SOC. These alerts were completely analyzed by the T2 analyst. The result is an information set with all information utilized by the analyst during their analysis.

Process construction. Having identified the steps of the analysis process, the T2 analyst mapped each step to the stages reported in Table 1. The T2 analyst then employed the obtained mapping to reconstruct the process they employed during their analyses.

Process verification. The obtained threat analysis process is then given in input to two separate senior analysts (one T3 analyst with 15+ years of experience and a security researcher with 10+ years of experience in threat analysis). Each expert is

asked to independently analyze three security alerts randomly obtained from the SOC environment (distinct from the ten employed for the process derivation) using the provided threat analysis process. Each senior analyst independently provided the T2 analyst with feedback on the process and considered information, and the process is updated accordingly. This process verification and update loop was repeated until all three experts agreed on the devised threat analysis process.

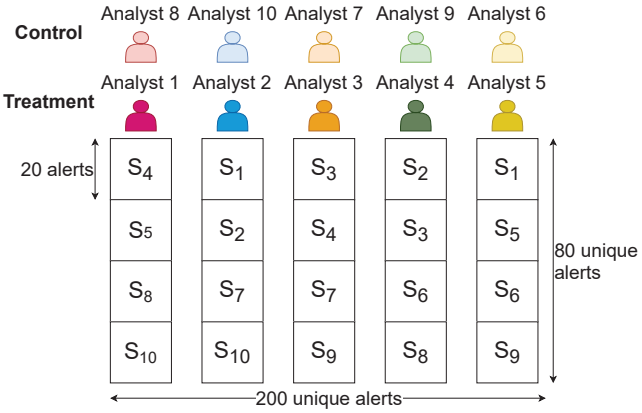
4.2 Experimental evaluation

To evaluate the effect of our threat analysis process on analysts' accuracy (**RQ2**), we ran an experiment involving T1 analysts and real alert data from one of the SOC sensors.

Experimental design. Figure 2 provides an overview of the experimental design. From the SOC environment we sampled 200 alerts and divided these in ten batches ('scenarios') of 20 alerts each. We then recruited ten T1 analysts and asked them to analyse four batches of alerts each (for a total 80 alert analyses per analyst);[3] each analyst was assigned to either the treatment group (i.e., following the proposed process for the analysis) or the control group, and asked to classify each alert as either 'interesting' or 'not interesting'. These assessments are then compared against a ground truth of assessments (defined by the SOC's T2 analyst) to evaluate differences in accuracy between the treatment and control groups. In the following, we describe our choice of subjects, how we designed the sets of alerts per subject, how we derived the ground truth and the details of the experimental setup.

Subjects. We recruited as subjects of our experiment ten junior analysts over a period of six months, in two batches of five analysts each. To maintain comparable experience levels across recruited analysts, we recruited them immediately after they joined the SOC. T1 analysts in the SOC are structurally hired as interns from the security program at a technical university in Europe. Their turnover rate varies between four and six months of employment.[4] All subjects are assessed before joining the team for their background, and are given the same technical on-the-job training. Analysts in the treatment group were given an additional training on the devised threat analysis process they will employ during the experiment.

Designing alert sets. *Collecting 'baseline' alert data.* To maintain realism of the experimental setup, we collected alerts from the network environment of a customer of the SOC over the course of 2.5 weeks. To make sure the collected alerts were not already investigated by our subjects as part of their normal job activities, we selected a network whose



Each analyst evaluates overall 80 alerts over four of ten scenarios. Each of the ten scenarios contains 20 unique alerts, at least one of which related to an injected attack, for a total of 200 unique alerts across scenarios. Each scenario is assigned to a single analyst at most once, and is analysed by two different analysts per experiment condition. For example, S_4 is analysed by Analyst 1 and Analyst 3, who are assigned to the treatment group, and Analyst 8 and Analyst 7, who are assigned to the control group.

Figure 2: Overview of experiment design

data is only captured for technical testing purposes by the SOC (as opposed to for security monitoring). The environment from which the alert data is collected comprises over 1500 unique hosts and multiple DNS and file servers. We logged approximately 100M connections attempts and 48M DNS requests. These connections generated 350k security events distributed across 150 unique security alerts. As SOC data are over-represented by alerts of certain kinds (e.g. alerts related to scan activities), we employed a stratified random sampling method over the collected alerts. We employ the 'rule category' [17] attribute that comes with alerts to define the type of each alert. The considered alert categories are: `Scan`, `Malware`, `CnC` and `Policy`. From each of the rule categories, we randomly selected a unique rule, and from that we sampled a random alert generated by that rule.

Generating 'successful' attacks. To generate 'interesting' alerts we could not rely on existing data in the SOC, as actual attacks are rare. We therefore employed PCAP network traffic from malware-traffic-analysis.net [15], which provides records of malicious network traffic of (multi-stage) malware attacks, to inject simulated attacks to the SOC sensors generating security alerts. Details of the attacks are reported in Appendix A. To assure the realism of the attacks, IP addresses in the obtained PCAPS are adapted to those expected within the range of the monitored network from which the 'baseline data' is derived. To avoid conflicts in the data, only unassigned IP addresses are used to rewrite the PCAPs; only internal IPs to the infected network were changed (i.e. IPs of the malware infrastructure remained unaltered). DNS servers used in the attacks are set to be the actual internal DNS servers in that sub-net, reflecting the corporate policy for the sub-net of the monitored environment. To inject the PCAPS in the SOC

[3] In consultation with the T2 and T3 analysts involved in this research, we estimated an average assessment time of 10 minutes per alert. Therefore, expected that no scenario would take more than 4 hours of a T1 analyst's time.

[4] The high turnover rate is due to the SOC's contractual policy (that follows the study program followed by the student analysts at the time of their recruitment), rather than to a high 'drop out' rate.

Table 2: Alert distribution across alert categories

Category	Ground truth	Overall
Command and Control	12	12
Malware	13	39
Policy	5	35
Scan	20	114
Total	50	200

network sensor to generate security alerts, we employed the SAIBERSOC tool [18].

Overall, our scenarios consist of 178 alerts sampled from the baseline monitored environment and 22 alerts generated by the injected attacks, for a total of 200 alerts. The 'Scenarios' ($S_1, S_2, \ldots, S_{10}$ in Figure 2) were then created by constructing 10 non-overlapping sets of 20 alerts. Each scenario contains alerts generated by exactly one injected attack; each attack generates at least one (and at most four) alerts.

Ground truth derivation. Once we derived our scenarios and related alerts, the T2 analyst ran a blind analysis over 5 alerts per scenario (for a total of $10 \times 5 = 50$ alerts) to label them as 'interesting' or 'not interesting'. All alerts related to an attack in a scenario were included in the set given to the T2 analyst. The remaining alert(s) for the ground truth were chosen randomly.[5] The distribution of alerts per category is reported in Table 2.

Experimental setup. The first batch of five analysts was assigned to the treatment condition. This choice was motivated by the need to empirically verify the internal consistency of the baseline threat analysis process before booking analysts' time away from the SOC. Details are reported in Appendix D. This batch received an in-depth training on the devised threat analysis process; the training was delivered by a T2 analyst, during the T1 analysts' intake at the SOC. The second batch was assigned to the control condition and only received generic training that was in place at the SOC before the introduction of the devised process. Analysts from both batches were asked to record their classification for each of the twenty alerts in a scenario as 'interesting' or 'not interesting', and to motivate their decision in plain English. Additionally, analysts from the first (treatment) batch were asked to record their evaluation for each of the steps identified in the threat analysis model for each of the analyzed alerts, in a separate worksheet.

4.3 Ethical considerations

This research was executed under ethical approval from our institution's ethical review board under approval number ERB2022MCS20. We gained explicit and informed consent from all subjects to participate in this experiment. Subject's names were anonymized to disassociate their identity from any performance evaluations. Furthermore, to alleviate the workload of our subjects, subjects participated in the experiment during working hours, as opposed to participating on top of their regular commitment with the SOC.

[5] Classifying the entire set of 200 alerts was not feasible due to the required time during which the T2 analyst would have been unavailable to the regular SOC operations.

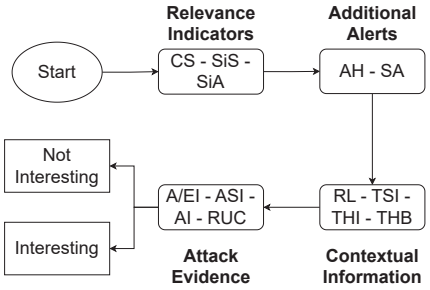


Figure 3: Overview of the baseline threat analysis process

5 A threat analysis process for network events

Following the iterative process described in Section 4.1, agreement on the details of the threat analysis process was reached at the fourth feedback iteration (at which point none of the three experts had any further remark). The result is 13 information steps mapped into the 4 stages reported in Table 1. Table 3 provides a summary of the final mapping between the process steps and stages. The final analysis process is visualized in Figure 3. The process guides the analyst in collecting evidence of an attack through the four identified stages; at the end of the process, the analyst decides whether there is enough evidence to classify the alert as 'interesting', or not.[6] The rest of this section details each step for every stage within the proposed threat analysis process.

5.1 Relevance Indicators

The first process stage consists of three steps; signature specificity (**SiS**), signature age (**SiA**) and customer scope (**CS**).

Signature specificity. To determine 'specificity', the analyst first determines whether the triggered signature is specific to a certain attack or service, or whether it is only a generic indicator of an attack. This step allows one to establish the initial priority of the alert analysis (e.g. specific indicators may be prioritised over generic indicators), as well as determine what to investigate in future steps. Generally, identifying that

[6] Importantly, we note that this process is not meant to be prescriptive, in that it does not provide instructions or thresholds to make specific decisions on the classification. Differently, it provides a framework of reference for the analyst to collect relevant information to make well-informed decisions on what action to take (i.e., ultimately, escalate or not escalate). Whether this decision can be at least partially automated or scripted away (on the basis of the collected information), or safely taken at a specific stage of the process, is out of the scope of this contribution.

Table 3: Mapping between the stages and the steps

Stage	Step	Description
Relevance indicators	signature specificity(**SiS**)	Indication of how specific the trigger condition is for the signature of the alert. (i.e Whether the signature is easily triggered).
	signature age(**SiA**)	The creation date of the signature, as new signatures are often more trustworthy and up-to-date than older ones.
	customer scope(**CS**)	Whether the alert is within the agreed scope of monitoring.
Additional alerts	alert history(**AH**)	The history of the same alert in the past. i.e. how commonly the alert triggerd in the past and for what reasons. This step is useful to detect common false positives.
	surrounding alerts(**SA**)	Other alerts relating to the specific alert under investigation. This step is useful for identifying alerts related to the same attack.
Contextual information	related logs(**RL**)	Logs related to the alert under investigation.
	traffic stream information(**TSI**)	The volume and content of the packets involved between the attacker and defender, compared to the expected volume and content for the protocol used.
	target host information(**THI**)	Any available information about the possibly affected host, such as whether it is a server or desktop, its OS etc.
	target hosts behaviour (**THB**)	The change in behavior of the host after the persumed attack
Attack evidence	attack/exploit information(**A/EI**)	The exact attack, objectives of the attack and the tools involved. The analyst can estimate the impact of the attack to its customer using this information.
	attacker information (**AI**)	Information about the attackers behavior, and whether the attack is from an unknown source.
	attack success indicators (**ASI**)	Information regarding whether the presumed attack was successful, such that the analyst may decide to not escalate unsuccessful attacks.
	relation to the use cases (**RUC**)	Indication of how much the possible attack overlaps with the use cases of the affected environment. The analyst considers the impact of the attack to the affected environment in this step.

the signature is specific here increases confidence in the event being interesting. Identifying that the signature is generic may decrease the confidence, depending on the level of generality.

Signature age. Analysts can check the signature creation date and last updated data of an alert to estimate if the behaviour triggering the alert is a recent or an old threat. This signals the age of associated threat, or indicates a potentially not interesting security event if the trigger conditions of the indicator are time-dependent.

Customer scope. The analyst determines if the alert is within the monitored scope by reviewing if necessary the customer security policy, as well as the service level agreements about sub-nets and reporting. The alerts with no to low impact, for example a guest network of the customer, can in this way be evaluated early on in the process as lower priority.

5.2 Additional alerts.

The second stage consists of two steps; alert history (**AH**) and surrounding alerts (**SA**).

Alert history. The analyst investigates the history of the alert under investigation. Namely, how often the alert has been triggered in the past, how often it was considered interesting, and whether it triggered for the same internal host before. Using this information, an analyst can verify quickly whether the observed alert is a common false positive or not. If past occurrences have been flagged as 'not interesting' due to them being false positives, the analyst can consider that the alert under investigation may be a false positive as well.

Surrounding alerts. The analyst investigates additional alerts similar to the one under investigation that were triggered by one or multiple of the involved hosts, around the time of the potential attack. When looking at these surrounding alerts, analysts may observe different alerts with similar names, indicating the same potential attack. This adds evidence that the event underlying these alerts may be interesting. Additionally the analyst may observe alerts for different phases of an attack, further strengthening the case for continuing to investigate the alert. For example, investigating a malware alert, a surrounding alert may be CnC activity. Identifying these surrounding alerts allows the analyst to get a more encompassing picture of an ongoing attack, if present.

5.3 Contextual information.

The third stage consists of four steps; related logs (**RL**), traffic stream information (**TSI**), target host information (**THI**) and target hosts behaviour (**THB**). At this stage, the analyst collects concrete evidence generated by the security systems, and information provided by the owner of the monitored environment. If the analyst finds no evidence of a potential attack reaching a vulnerable host, the analyst may consider it as evidence to classify the alert as 'not interesting'.

Related logs. This step focuses on identifying logs useful

to evaluate the cause or the outcome of the attack under investigation. This selection is largely dependent on the type of alert, and the type of traffic it is triggered on. In general, **RL** consists of at least a connection log, a protocol specific log (such as HTTP, or SSH), in addition to the alert log. Furthermore, any other logs generated by the receiving host of the protocol in the alert, or DNS logs, are typically related.

Traffic stream information. The analyst considers more detailed information about the packets sent to and from the host. This information can include the total number of bytes and packets sent by the attacker and defender, and the data contained in those packets. The protocol used between the communication of the attacker and defender is an important consideration in this step as it determines whether the number of packets and the data contained in them are abnormal or not in the specific context. For example, this step allows analysts to identify successful port scans by verifying if a response packet was sent back to the source. This also allows analysts to determine whether any 'lucky hits', were generated. A 'lucky hit' occurs when the trigger conditions of a signature (typically a non-specific one, as assessed in the **SiS** step) are met by pure chance on a random sequence of bytes, and thus trigger on benign traffic, producing false-positives.

Target host information. The analyst can utilize information about the host, such as whether it is a desktop or a server, its purpose (for example DNS server), its OS, its associated sub-net, host name, open ports and so on, to reason about whether the attack under investigation can ever lead to successful violation of corporate policies. This step can vary greatly from SOC to SOC and even from monitored environment to monitored environment, as corporate policies differ between organizations.

Target host behaviour. Next to utilizing known information about the targeted host, the analyst reviews the current behavior of the targeted host. For this, the logs produced by the IDS and network sniffer are utilized to review the behaviour of the host before and after the attack. If the host behaves abnormally compared to how the hosts normally would behave, it may indicate that the host was impacted by the attack.

5.4 Attack evidence

This stage consists of four steps; attack/exploit information (**A/EI**), attack success indicators (**ASI**), attacker information (**AI**) and relation to the use cases (**RUC**).

Attack/Exploit information. In this step, the analyst determines the exact attack and tools involved. The information required to determine this originates from the signature which triggered the alert, and open source information about the corresponding attack. From this step, it should be clear whether there is an attack, and if so what attack specifically. Using this information, in relation to that collected in the previous steps (e.g. **RL**, **THI**) the analyst estimates how this specific attack can have impact on the customer.

Attacker information. The analyst investigates the behaviour of the attacker (or at least of the attacking system). Using the logs generated as a result of the attacker behaviour, as well as using public sources, the analyst can determine whether the attacker is an actual attacker. This step is needed to rule out known and trusted sources such as (vulnerability) scanners, as well as help identifying false positive alerts generating 'hits' on backup streams, software updates, and benign network downloads.

Attack success indicators. The analyst investigates whether the attack was successful. Analysts use information obtained from former stages and open sources that identify clear indicators of successful attacks. Generally, the attack success indicators are highly dependent on the specific attack, however, generic indicators such as DNS requests for unusual top-level domains or internal scanning can be used as well.

Relation to use cases. The analyst consults the use cases for the affected environment. This step helps them to correctly identify the full impact for the environment, and thus the final classification of the alert. It also eliminates any alerts which are not important to the environment. For example, investigating a generic malware alert, having determined that it is actually adware on a desktop, the use cases may call for no action at all, depending on the environment. Finally, the use cases may provide useful information and guidance on what to report to higher tier analysts or the affected customer.

6 Experiment results

Table 4 provides an overview of the alert analyses performed by our subjects. Collectively, analysts classified an alert as 'interesting' 114 times, and 686 times as 'not interesting'. Furthermore, we observe that analysts who followed our process classify alerts more often as 'interesting' (67 times) than the analysts who did not follow the process (47 times, $\chi = 3.69, p = 0.055$). Whereas only borderline significant, this suggests that following the proposed process may increase the likelihood of escalating an alert to a higher tier. Meanwhile, we do not observe any within-group difference across analysts in terms of their classification outputs in either group (treatment: $\chi = 1.36, p = 0.85$; control: $\chi = 2.07, p = 0.72$). This suggests that the likelihood of an analyst's classification for a given alert may depend on the treatment group the analyst is assigned to, rather than on the analyst themselves.

Focusing on analysts accuracy, we observe that overall analysts not following our process show a accuracy of 82% in the classification; by contrast, analysts following the proposed process show an overall accuracy of 92%. Interestingly, this difference disappears when only considering alerts whose ground truth classification is 'not interesting'. By contrast, 'interesting' alerts were classified correctly only 65.9% (29 our of 44 possible assessments on 'interesting' alerts') of the times by analysts not employing our process, while the group

Table 4: Overview of analysts' classifications

		All alerts		Alerts included in ground truth					
				Interesting		Not Interesting		Total	
Analyst	Process	Int.	Not Int.	Correct	Wrong	Correct	Wrong	Correct	Wrong
1	Yes	14	66	8 (88.9%)	1 (11.1%)	11 (100.0%)	0 (0.0%)	19 (95.0%)	1 (5.0%)
2	Yes	10	70	7 (87.5%)	1 (12.5%)	11 (91.7%)	1 (8.3%)	18 (90.0%)	2 (10.0%)
3	Yes	14	66	7 (87.5%)	1 (12.5%)	11 (91.7%)	1 (8.3%)	18 (90.0%)	2 (10.0%)
4	Yes	15	65	8 (100.0%)	0 (0.0%)	11 (91.7%)	1 (8.3%)	19 (95.0%)	1 (5.0%)
5	Yes	14	66	9 (81.8%)	2 (18.2%)	9 (100.0%)	0 (0.0%)	18 (90.0%)	2 (10.0%)
6	No	9	71	9 (81.8%)	2 (18.2%)	9 (100.0%)	0 (0.0%)	18 (90.0%)	2 (10.0%)
7	No	8	72	4 (50.0%)	4 (50.0%)	11 (91.7%)	1 (8.3%)	15 (75.0%)	5 (25.0%)
8	No	11	69	8 (88.9%)	1 (11.1%)	10 (90.9%)	1 (9.1%)	18 (90.0%)	2 (10.0%)
9	No	7	73	5 (62.5%)	3 (37.5%)	12 (100.0%)	0 (0.0%)	17 (85.0%)	3 (15.0%)
10	No	12	68	3 (37.5%)	5 (62.5%)	11 (91.7%)	1 (8.3%)	14 (70.0%)	6 (30.0%)
				Overall					
With process		67	333	39 (88.6%)	5 (11.4%)	53 (94.6%)	3 (5.4%)	92 (92%)	8 (8%)
Without process		47	353	29 (65.9%)	15 (34.1%)	53 (94.6%)	3 (5.4%)	82 (82%)	18 (18%)
Total		114	686	68 (77.2%)	20 (22.8%)	106 (94.6%)	6 (5.4%)	174 (87%)	26 (13%)

who did follow the process classified the same set correctly 88.6% (39/44) of the times. This suggests that the proposed process is particularly useful for alerts related to attacks, reducing the classification inaccuracy by more than 20%. Generally, we find T1 analysts to perform better at classifying 'not interesting' alerts as opposed to 'interesting' alerts with a classification accuracy of 94.6% and 77.2% respectively.

To evaluate the effects of the proposed threat analysis process on assessment accuracy, we perform a logistic regression on the dependent variable *Correct*, which is a dummy variable set to 1 if an analyst correctly classifies the security alert, and 0 otherwise. The explanatory variables in the regression model are *Process* and *Category*. *Process* is a dummy variable set to 1 if the analyst followed our threat analysis process; *Category* is a categorical variable representing the category of an alert among the categories `Scan`, `Malware`, `CnC` and `Policy`. In addition, we run checks to evaluate whether a mixed effect model is required to account for the fact that multiple observations are assessed per subject, and checks to account for additional effects caused by the specific scenarios. To do this, we consider whether `Analysts` or the `Scenarios` play a role in the outcome. We run two separate logistic regression models: one with analyst dummy-variables as predictors, and one with scenario dummy-variables as predictors. Table 5 provides an overview of the results. For both models, we find no significant effect of any analyst or scenario on the predicted outcome. A joint ANOVA test confirms this as the null-hypothesis of all coefficients being equal to zero is not rejected, which is consistent with `Analyst` and `Scenario` not playing a role in differentiating assessments ($p = 0.365$ and $p = 0.323$ respectively). We therefore do not include either variable in the final model presented here, and use logistic regression to fit the model $Correct = c + \beta_1 Process + \beta \mathbf{Category}$.[7]

Table 5: Logistic regression on the correctness of evaluations: once with subject dummies and once with scenarios dummies.

Variable	Coeff.	p	Variable	Coeff.	p
(Intercept)	2.94	0.004	(Intercept)	2.20	0.003
Analyst 2	-0.75	0.556	*Scenario* 2	-0.46	0.635
Analyst 3	-0.75	0.556	*Scenario* 3	-0.81	0.384
Analyst 4	<0.01	1.000	*Scenario* 4	-0.81	0.384
Analyst 5	-0.75	0.556	*Scenario* 5	<0.01	1.000
Analyst 6	-0.75	0.556	*Scenario* 6	0.75	0.556
Analyst 7	-1.85	0.108	*Scenario* 7	-1.35	0.130
Analyst 8	-0.75	0.556	*Scenario* 8	0.75	0.556
Analyst 9	-1.21	0.314	*Scenario* 9	0.75	0.556
Analyst 10	-2.10	0.065	*Scenario* 10	<0.01	1.000

Table 6: Logistic regression on the correctness of evaluations, as dependent on the used process and the alert category

Variable	Coeff.	OR change (%)	p-value
(Intercept)	2.56	NA	<0.001
Process	**0.98**	**167.0**	**0.035**
Reference category: `Scan`			
Category : `CnC`	-1.20	-69.8	0.070
Category : `Malware`	**-1.89**	**-84.9**	**0.002**
Category : `Policy`	-0.76	-53.1	0.406

Table 6 shows the effect sizes alongside associated p-values from the fixed effects logistic regression model. Coefficients shown in bold denote an associated p-value of 0.05 or less which we consider statistically significant. As `Scan` is the most common alert category in the SOC, we choose it as the

[7] For completeness, we also estimated the mixed effects model. The coefficients are qualitatively identical both in magnitude and direction to those reported here.

baseline category for the variable `Category`; coefficients for other categories should therefore be interpreted relative to it. The coefficient of *Process* is 0.98 with a p-value of 0.035, showing that analysts following the proposed threat analysis process were significantly more likely to classify alerts correctly than analysts in the control group. This corresponds to a change in the odds of correct classification of 167%, i.e. a shift in probability of generating a correct assessment from approximately 82% in the control group to 92% in the treatment group. We also observe that there is a significant difference between `Malware` and `Scan` ($p = 0.002$) alerts. Malware related alerts were more often incorrectly assessed than scan alerts, indicating that Malware alerts are significantly more difficult to analyze correctly. Other differences across categories are smaller and not statistically significant.

6.1 Qualitative evaluation

We now try to qualitatively characterize the differences between the two groups by looking at specific classification tasks in the two groups. To reconstruct this, we look at the data annotated by the analysts with the motivations of their decisions for a classification for each specific alert.

Firstly, from the data we observed that subjects who do not follow our process typically based their decision to discard an alert (i.e classifying it as 'not interesting') after a single 'step' in the decision process. For example, a `CnC` alert (`ThreatFox BazarBackdoor botnet C2 traffic`, whose instance in our data is classified as 'interesting' by the T2 analyst) in scenario no. 3 was erroneously classified by an analyst as 'not interesting' as the network communication related to this alert "only" contained 10 packets. One of the analysts remarks:"*Its [the count is] below 50. So, this alert can also be dismissed.*" Whereas '50' is not a limit specified anywhere in the SOC for this type of alert, we later learned that this cut-off number is considered relevant for SSH brute force attacks. This suggests that the analyst erroneously considered this a universal threshold when deciding whether a communication is large enough to be considered potentially interesting, despite the alert in question being completely unrelated to SSH brute forcing. This seems in line with the generally accepted notion of 'implicit knowledge' being employed by analysts [5, 22]. Although our process does not prevent these mistakes from happening, following it may at least aid analysts in considering other steps as well to potentially classify alerts in a more informed manner.

In another investigation on alert `ET JA3 Hash - [Abuse.ch] Possible Dridex`, two analysts who erroneously classified it as 'not interesting' had previously observed high false positive rates with alerts associated to 'JA3' hashes. Therefore, analysts investigating JA3 alerts often mumbled that this is most likely going to be a false positive. Further, this alert was associated to a limited (9) number of packets, leading analysts not following the process to classify it as 'not interesting' despite the presence of concrete evidence of a connection from a suspicious IP being established with the host. By contrast, analysts in the treatment group identified that this alert was related to another 'interesting' alert at the **SA** step. Whereas the T1 analysts could not observe much of the data relating to the network communication of this alert, they identified sufficient evidence to escalate it, considering that if the alert was related to an attack it would have a high impact to the organization. One of the analysts following our process remarked the following related to this alert: "*Could not tell that the decoded message would have had a relation to this traffic. It is still malware-related, making it more significant for the customer and this same IP was also involved with the Threatfox backdoor alert.*" Interestingly, another analyst following the process and correctly classifying this alert as 'interesting' commented "*It's weird. JA3 is never interesting for us.*". This suggests they made similar considerations to the analysts not employing the process, but corrected their belief on the basis of the additional evidence collected.

We find three cases where following the process lead to analysts classifying a 'not interesting' alert as 'interesting', i.e. generating a false positive classification. One scan alert (`ET SCAN MS Terminal Server Traffic on Non-standard Port`) was classified as 'interesting' because there was insufficient evidence in one 'step' of the investigation. Almost all 'steps' in this investigation were leading to the conclusion that the alert was indeed not interesting. However, as the subject could not observe the behavior of the host, the subject decided to classify it as 'interesting' nonetheless to verify the alert with a T2 analyst. Another interesting case was when an analyst over-relied on **AI** instead of other steps in the process when investigating `ET SCAN ProxyReconBot CONNECT method to Mail`. This alert was raised despite the attempted scan receiving no response packets from the host. Yet, as the IP which was scanning the network corresponded to an untrusted domain, the T1 analyst decided to classify the alert as 'interesting'. Overall, these errors seem to be caused by a mistaken interpretation of the evidence (or lack thereof) by the analyst, rather than being induced by an incorrect evaluation strategy imposed by the process.

7 Discussion

Our findings show that analysts are significantly more likely to classify alerts correctly (odds increase by around 2.5 times) when following our baseline threat analysis process. This suggests that a structured process that T1 analysts can follow can when compared to sole reliance on tacit knowledge [5], aid in the correctness of security alert classification. Interestingly, in our experiment this increase in accuracy can be mainly attributed to 'interesting' alerts. In our experiment, we observe that for our analysts the rate of correct classifi-

cations of a 'not interesting' alert is higher than 90%; this suggests that a 'not interesting' alert may be easier to analyse and thus may benefit to a lesser extent from a structured way of processing information. For example, 'not interesting' alerts raised by attempted (but failed) port scans can often be dismissed by simply observing that the host system did not communicate back to the attacker. Therefore, most analysts would classify such alerts correctly, regardless of how rigorously they analyse the evidence. By contrast, T1 analysts in our experiment struggled more with analysing 'interesting' alerts correctly. This is unsurprising as these alerts are in general more complex and require the analysis of more information. Considering a delta of 20% in correct assessments for 'interesting' alerts between the two experimental conditions, our results suggest that structuring the analysis process may improve the classification accuracy specifically for the hardest alerts to analyze. Our example in Section 6.1 illustrates that this may be the case as analysts who do not follow our process may over-rely on one information point and simultaneously not consider other relevant information required for the analysis, whereas analysts who do eventually find the relevant information.

7.1 Implications for practice

Training. SOCs conduct training for their T1 analysts to for example, update analysts on the latest threats and how to analyze them [8, 11, 14]. As the training directly improves the effectiveness of analysts, it is considered a crucial aspect of a SOC [2, 11, 20]. However, T1 analysts need to have a baseline level in their work, such that analysts are able to perform adequately even if they have not been trained on that specific set of alerts. By structuring the workflow of a T1 analyst, it streamlines the baseline knowledge a T1 analyst should have in a SOC. The specific information T1 analysts should collect is explicitly defined, and thus SOCs can tailor their training towards how to collect the required information.

Measuring analyst performance. It is important for SOCs to measure the performance of their analysts such that they know where different detection tools or more training are required. For example, if a SOC realizes that analysts are having significant difficulties interpreting relevant logs, it may consider training their analysts on logs specifically. However, current quantitative metrics for SOC analysts often fail to measure the actual performance of the analyst [14, 22]. Our proposed threat analysis process can standardize the workflow of T1 analysts in terms of what information they should collect during their analyses. This gives SOC managers more concrete directions to measure the performance of their analysts, and of the processes they oversee. Although, it is out of the scope of this paper to present better metrics for T1 analyst performance, measuring performance of analysts at specific steps gives a more accurate overview of their analysis performance as opposed to only considering the number of escalated alerts [14], handled alerts [22] and time needed to analyse an alert [14]. Importantly, this may reveal 'weak' spots in the detection and escalation processes in place at a SOC, giving managers accurate metrics on which to base future adjustments.

Escalation. When an alert is being escalated by a T1 analyst, the analyst escalates the alert itself with supporting evidence why the alert has been escalated [13]. However, what may constitute as supporting evidence may differ for each individual analyst. SOCs may have their own standards and expectations on what T1 analysts include in their 'ticket'. On the other hand, the proposed threat analysis process (or any structured process analysts can follow) can be used to provide a ticketing standard that is in tune with the process that the T1 analyst follow. Furthermore, by removing uncertainty on the expectations of a T2 analyst on what information they will receive from a T1 analyst, the time needed to interpret each ticket by a T2 analyst may be reduced.

7.2 Implications for research and future work

There have been numerous previous studies presenting proposed tools pertaining to issues in the threat analysis process of SOC analysts [3, 4, 9, 31]. In line with this, future work could integrate a threat analysis process into an operational SOC. In our work, a subset of analysts were required to follow our process, however this is hard to enforce outside the controlled experiment. Observing how analysts would classify real security events using an integrated system that guides them into a desired process would potentially yield interesting insights into how such a process would function in practice.

Similarly, future work may evaluate 'how much information' is 'enough information' to collect to take an accurate decision on a specific alert. This may aid the navigation of an analysis process for analysts to 'quit' the process early on when enough evidence has been collected to take a negative decision. Similarly, the proposed process may be extended to other types of data, e.g. considering host or cloud log data rather than network (event) data.

In our work we focused on the accuracy of the classification of an analyst as the sole metric of a T1 analyst. However, previous studies have shown that timeliness of analysis is important as well in an operational SOC [8, 22]. Even if our threat analysis process leads to a better outcome in classification, it would be problematic if it added a significant time overhead for T1 analysts. Future work could investigate how following such a process influences the timeliness (and not only the accuracy) of the classification of security events.

7.3 Threats to validity

Construct validity. All injected attacks in our experiment consisted of malware related attacks, where a malware is installed, the host is controlled via a command and control server and where possibly some lateral movement took place

within the network. Considering that SOCs encounter other forms of attacks, a set of alerts generated by malware related attacks may not fully reflect the concept of 'interesting' alerts.

Internal validity. We assume in our experiment that all our subjects are equally skilled in analyzing security events and do not influence the accuracy of the classifications. However, in reality some analysts may be more skilled than others despite similar job experiences and educational background. To mitigate this, we checked whether concrete evidence exists that analysts influence the classification of the alerts or not. Additionally, when collecting data regarding the internal consistency of our process we used the response options in Table 7 in the Appendix. However, we did not test whether the interpretations of the response options differ among our subjects. In other words, subjects may give different responses to a step with the same observed data. Meanwhile, subjects may give identical responses to a step even though they have interpreted the data completely differently.

External validity. The main threat to external validity is the extent to which the employed SOC represents data and operations adopted by other SOCs. SOCs vary widely over both dimensions as they employ different technologies and sensors, SOC operations are not standardized, and by monitoring different networks/infrastructures they may evaluate different alerts over different environments [25]. However, virtually all SOCs at least monitor network traffic [6, 16, 27] and typically employ junior T1 analysts as a first line of defense to decide whether to escalate or ignore incoming alerts [25]. As the SOC under analysis performs only network analysis, and employs T1 analysts from the same 'pool' as most SOCs (i.e., junior staff in need of specialized cybersecurity training to operate well within a SOC [25]) we consider it to be representative of SOCs in general, over these two dimensions. Further, as the SOC under analysis *only* performs network monitoring, we can evaluate model effects without additional confounding factors caused by multiple data sources (e.g. system host logs). Whereas this suggests that our finding that a structured analysis process can help analysts in making accurate evaluations over network alerts, effect sizes may vary significantly across SOCs. Further research is needed to derive and evaluate analysis processes across different SOCs, monitored environments, and monitoring technologies, and their interactions. Additionally, whereas the collaborating SOC only allows two possible classifications for T1 analysts, past research [21] show that other SOCs may have more options to classify an alert. Our threat analysis process does not incorporate such frameworks for classifying alerts and thus, our process requires modification to accommodate different classification systems.

8 Conclusions

In this work we devised a threat analysis process to attempt to structure the work process of T1 SOC analysts. Our threat analysis process consists of four stages where it guides the analyst into collecting information relevant for their analysis. Furthermore, we conducted an experiment using real alert data with ten T1 analysts working in a commercial SOC to investigate the effect of structuring the threat analysis process. Our results show that our process increases the odds of our subjects correctly classifying an alert by 167%. More specifically, we observed that alerts correlated with a cyber attack are the alerts who significantly benefit from using our threat analysis process. Overall, our study suggests that structuring the analysis process of a T1 analyst aid in the correct classification of security alerts.

Acknowledgments

This work is supported by the SeReNity project, Grant No. cs.010, funded by Netherlands Organisation for Scientific Research (NWO) and by the INTERSECT project, Grant No. NWA.1162.18.301, funded by NWO. The authors also thank the Eindhoven security hub SOC for its collaboration in this work.

References

[1] Alan Agresti. *Categorical data analysis*, volume 792. John Wiley & Sons, 2012.

[2] Bushra A. Alahmadi, Louise Axon, and Ivan Martinovic. 99% false positives: A qualitative study of SOC analysts' perspectives on security alarms. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 2783–2800, Boston, MA, August 2022. USENIX Association.

[3] Kenneth B. Alperin, Allan B. Wollaber, and Steven R. Gomez. Improving interpretability for cyber vulnerability assessment using focus and context visualizations. In *2020 IEEE Symposium on Visualization for Cyber Security (VizSec)*, pages 30–39, 2020.

[4] Louise M. Axon, Bushra A. AlAhmadi, Jason R. C. Nurse, Michael Goldsmith, and Sadie Creese. Sonification in security operations centres: what do security practitioners think? *CoRR*, abs/1807.06706, 2018.

[5] Selina Y. Cho, Jassim Happa, and Sadie Creese. Capturing tacit knowledge in security operation centers. *IEEE Access*, 8:42021–42041, 2020.

[6] A. D'Amico and K. Whitley. *The Real Work of Computer Network Defense Analysts*, pages 19–37. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[7] Anita D'Amico, Kirsten Whitley, Daniel Tesone, Brianne O'Brien, and Emilie Roth. Achieving cyber defense situational awareness: A cognitive task analysis

of information assurance analysts. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 49, pages 229–233, 09 2005.

[8] Varun Dutt, Young-Suk Ahn, and Cleotilde Gonzalez. Cyber situation awareness: Modeling the security analyst in a cyber-attack scenario through instance-based learning. In *20th Annual Conference on Behavior Representation in Modeling and Simulation 2011, BRiMS 2011*, pages 280–292, 07 2011.

[9] Roman Graf, Florian Skopik, and Kenny Whitebloom. A decision support model for situational awareness in national cyber operations centers. In *2016 International Conference On Cyber Situational Awareness, Data Analytics And Assessment (CyberSA)*, pages 1–6, 2016.

[10] Eric T. Greenlee, Gregory J. Funke, Joel S. Warm, Ben D. Sawyer, Victor S. Finomore, Vince F. Mancuso, Matthew E. Funke, and Gerald Matthews. Stress and workload profiles of network analysis: Not all tasks are created equal. In Denise Nicholson, editor, *Advances in Human Factors in Cybersecurity*, pages 153–166, Cham, 2016. Springer International Publishing.

[11] Robert Gutzwiller, Sunny Fugate, Ben Sawyer, and Peter Hancock. The human factors of cyber network defense. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 59:322–326, 09 2015.

[12] Wajih Hassan, Shengjian Guo, Ding Li, Zhengzhang Chen, Kangkook Jee, Zhichun Li, and Adam Bates. Nodoze: Combatting threat alert fatigue with automated provenance triage. In *NDSS Symposium*, 01 2019.

[13] Christopher Healey, Lihua Hao, and Steve Hutchinson. Visualizations and analysts. *Advances in Information Security*, 62:145–165, 10 2014.

[14] Faris Bugra Kokulu, Ananta Soneji, Tiffany Bao, Yan Shoshitaishvili, Ziming Zhao, Adam Doupé, and Gail-Joon Ahn. Matched and mismatched socs: A qualitative study on security operations center issues. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, CCS '19, page 1955–1970, New York, NY, USA, 2019. Association for Computing Machinery.

[15] @malware_traffic. My Technical Blog Posts. https://www.malware-traffic-analysis.net/. [Online; accessed May 5, 2022].

[16] Joseph Muniz, Gary McIntyre, and Nadhem AlFardan. *Security operations center: Building, operating, and maintaining your SOC*. Cisco Press, 2015.

[17] Proofpoint. TECH BRIEFET Category Descriptions. https://tools.emergingthreats.net. [Online; accessed August 3, 2022].

[18] Martin Rosso, Michele Campobasso, Ganduulga Gankhuyag, and Luca Allodi. Saibersoc: Synthetic attack injection to benchmark and evaluate the performance of security operation centers. In *Annual Computer Security Applications Conference*, ACSAC '20, page 141–153, New York, NY, USA, 2020. Association for Computing Machinery.

[19] Reza Sadoddin and Ali Ghorbani. Alert correlation survey: Framework and techniques. PST '06, New York, NY, USA, 2006. Association for Computing Machinery.

[20] Sathya Chandran Sundaramurthy, Alexandru G. Bardas, Jacob Case, Xinming Ou, Michael Wesch, John McHugh, and S. Raj Rajagopalan. A human capital model for mitigating security analyst burnout. In *Eleventh Symposium On Usable Privacy and Security (SOUPS 2015)*, pages 347–359, Ottawa, July 2015. USENIX Association.

[21] Sathya Chandran Sundaramurthy, Jacob Case, Tony Truong, Loai Zomlot, and Marcel Hoffmann. A tale of three security operation centers. In *Proceedings of the ACM Conference on Computer and Communications Security*, 11 2014.

[22] Sathya Chandran Sundaramurthy, John McHugh, Xinming Ou, Michael Wesch, Alexandru G. Bardas, and S. Raj Rajagopalan. Turning contradictions into innovations or: How we learned to stop whining and improve security operations. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*, pages 237–251, Denver, CO, June 2016. USENIX Association.

[23] Sathya Chandran Sundaramurthy, John McHugh, Xinming Simon Ou, S. Raj Rajagopalan, and Michael Wesch. An anthropological approach to studying csirts. *IEEE Security & Privacy*, 12(5):52–60, 2014.

[24] Thijs van Ede, Hojjat Aghakhani, Noah Spahn, Riccardo Bortolameotti, Marco Cova, Andrea Continella, Maarten van Steen, Andreas Peter, Christopher Kruegel, and Giovanni Vigna. Deepcase: Semi-supervised contextual analysis of security events. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 522–539, 2022.

[25] Manfred Vielberth, Fabian Böhm, Ines Fichtinger, and Günther Pernul. Security operations center: A systematic study and open challenges. *IEEE Access*, 8:227756–227779, 2020.

[26] John Yen, Robert Erbacher, Chen Zhong, and Peng Liu. Cognitive process. *Advances in Information Security*, 62:119–144, 10 2014.

[27] Chen Zhong, Awny Alnusair, Brandon Sayger, Aaron Troxell, and Jun Yao. Aoh-map: A mind mapping system for supporting collaborative cyber security analysis. In *2019 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA)*, pages 74–80, 2019.

[28] Chen Zhong, John Yen, Peng Liu, Rob Erbacher, Renee Etoty, and Christopher Garneau. An integrated computer-aided cognitive task analysis method for tracing cyber-attack analysis processes. In *Proceedings of the 2015 Symposium and Bootcamp on the Science of Security*, HotSoS '15, New York, NY, USA, 2015. Association for Computing Machinery.

[29] Chen Zhong, John Yen, Peng Liu, Rob F. Erbacher, Christopher Garneau, and Bo Chen. *Studying Analysts' Data Triage Operations in Cyber Defense Situational Analysis*, pages 128–169. Springer International Publishing, Cham, 2017.

[30] Chen Zhong, John Yen, Peng Liu, and Robert Erbacher. Learning from experts' experience: Toward automated cyber security data triage. *IEEE Systems Journal*, PP:1–12, 05 2018.

[31] Chen Zhong, John Yen, Peng Liu, and Robert F. Erbacher. Automate cybersecurity data triage by leveraging human analysts' cognitive process. In *2016 IEEE 2nd International Conference on Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing (HPSC), and IEEE International Conference on Intelligent Data and Security (IDS)*, pages 357–363, 2016.

A Injected attacks

The list below identifies the attacks that were injected as part of the experiment, and the general behaviour that could be determined from the logs and alerts the attacks generated in the experiment environment. All attacks involve a malware(s). Alerts are generated from installations of such malware, command and control traffic or lateral movements. The table below shows which of the three aforementioned components of an attack generated an alert. **I** stands for installation, **CnC** for command and control and **LM** for lateral movements.

ID	Attack	I	CnC	LM
1	Remcos RAT	X	X	X
2	RIG Exploit Kit and Dridex	X	X	
3	Emotet and Trickbot		X	X
4	Qakbot and Cobalt Strike	X	X	
5	Qakbot and Spambot	X	X	
6	Hancitor and Cobalt Strike	X	X	X
7	Ghost RAT		X	
8	BazaarLoader and Cobalt Strike	X	X	X
9	MalSpam Brazil	X	X	
10	Ursnif	X	X	

B Environment

To ensure that the only additional training required for the experiment is the training related to our threat analysis process, we replicated the SOC environment on which our subjects work, and received their generic intake training. The environment is based on the Elastic Stack (ELK) and employs instrumented Suricata and Zeek sensors for the network event analysis (Suricata for attack detection and Zeek for logging network traffic). In our experiment Suricata was deployed with the open source *Emerging Threat Open ruleset*, as well as the licensed *Emerging Threat PRO ruleset* employed at the SOC. Replicating the configuration used in the production environment of the SOC, a subset of rules was configured to not trigger alerts (i.e., starting points for analyst investigations), but rather to generate logs stored in the SIEM. These logs can be used by analysts to further enrich the context of the events that triggered the investigated alert. These 'muted' signatures include *hunting*, *policy*, and *info signatures*. On top of the alerts and network logs generated by the sensors, the analysts were allowed to seek additional information from online sources (e.g. to check file hashes, IP address reputation, perform `whois` queries, ..) as normally performed during real operations. Because of storage limitations in the experiment environment, analysts did not have access to raw network traffic in PCAP files.

C Analysis sheet

Table 7 provides a summary of the options given to analysts for each of the process stages and steps identified in Fig. 3.

D Alert assessment consistency

We first evaluate whether the proposed threat analysis process produces consistent evaluations by the analysts. To evaluate this, we compare the evaluations made by analysts in the first batch across all steps of the proposed process. To do this, we compute the agreement score between analyses within the same scenario. The agreement score is calculated by counting the frequency per step where the two analysts outputted identical answers and then dividing it by the total number of alert instances (i.e 200).

We first consider the extent to which analysts agree on their assessments for each step of the process delineated in Table 3. Agreement scores for each step in the process are calculated across 200 pairwise comparisons of assessments performed by two separate analysts. Figure 4 shows the calculated agreement rates across each process stage and step. We find that, overall, analysts agree on the evaluation of the information

Table 7: Response option for each step

Stage	Step	No. Response options	Response options
Relevance indicators	signature specificity(**SiS**)	2	Old, New
	signature age(**SiA**)	2	Generic, Specific
	customer scope(**CS**)	2	Yes, No
Additional alerts	alert history(**AH**)	4	First occurence, Typically NI, Typically FP, Inconclusive
	surrounding alerts(**SA**)	2	Adds Evidence, Does not add evidence
Contextual information	related logs(**RL**)	4	Logs which indicate the result/impact of the event causing the alert, Logs which indicate the event which is the cause of alert, Both, None
	traffic stream information(**TSI**)	3	Small, Normal, Large
	target host information(**THI**)	2	Vulnerable, Not vulnerable
	target hosts behaviour (**THB**)	2	Normal behavior, Unusual behavior
Attack evidence	attack/exploit information(**A/EI**)	2	Attack, No attack
	attacker information (**AI**)	2	Trusted external host, Unknown external host
	attack success indicators (**ASI**)	2	Definitely unsuccessful, Successful, Unknown
	relation to the use cases (**RUC**)	2	Unrelated, Related

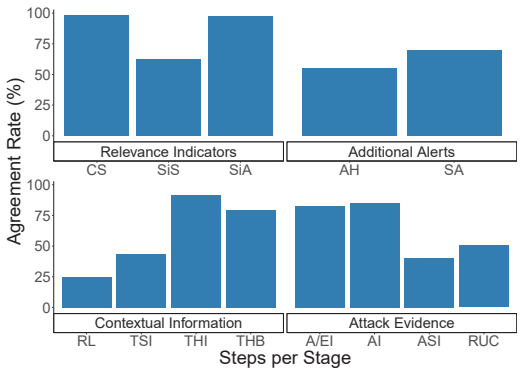


Figure 4: The internal consistency of our baseline threat analysis process.

collected in each step. We stress that the process does not mandate or instruct the analysts in how to find relevant information to make an assessment on that specific step (e.g. no query template is provided to identify 'related logs', `RL`, in the `Contextual information` stage). It is therefore to be expected that, the wider the information space associated to the assessment of a specific step is, the lower the expected agreement of analysts is. Our findings suggest that this holds also for our pool of analysts who have similar background and level of experience, and who received the same professional training. On the other hand, we observe that for each stage one or more steps consistently achieve relatively high agreement levels of 70% or more. To evaluate analysts agreement on the outcome of the process, we calculate Cohen's Kappa on analysts' final classification of an alert as 'interesting' or 'not interesting' for those analysts who employed the proposed process ($\kappa = 0.52$, $CI : [0.36, 0.68]$), and those who did not ($\kappa = 0.45,$, $CI : [0.28, 0.64]$). Whereas a straightforward interpretation of κ is not possible, both scores indicates a 'moderately strong agreement' [1, Ch.11.5.4] within the two groups. However, we do not find significant differences in the agreement levels between the two groups. This suggests that analysts in either group take similar decisions when compared to analysts in the same group.

E Changes to the process as a result of expert feedback

The initial sequence steps identified was extended with **CS** after the first round of verification, and **RUC** was moved from the start of the sequence to the end. **CS** was previously covered by **RUC**, but was found to be atomic and impactful enough to justify its own step. Additionally, **CS** can be determined more easily and thus earlier, than **RUC**.

As discovered during the verification by the experts, **RUC** requires details about the attack, the affected system and the impact, which are not available early on in the analysis process. For this reason as well, **RUC** was moved to the end of the sequence of steps. The adjustments detailed above were implemented before the experiment design and execution.

The question corresponding to the stage `Contextual information` was changed from "2-way communication established between attacker and attacked host" to "Vulnerable host reached by potential attack", to capture cases were attacks or exploits do not result in a two-way communication between the attacker and the attacked host.

Finally, a step named "signature quality" (now omitted) was split up into "signature specificity" and "signature age", the step "target host information" was split up into "target host information" and "target host behaviour" and the step "attack/exploit information" was split into "attack/exploit information" and "attacker information". These changes were made to make the steps more atomic.

Exploring the Security Culture of Operational Technology (OT) Organisations: The Role of External Consultancy in Overcoming Organisational Barriers

Stefanos Evripidou, *University College London* Uchenna D Ani, *University of Keele*
Stephen Hailes, *University College London* Jeremy D McK. Watson, *University College London*

Abstract

Operational Technology (OT) refers to systems that control and monitor industrial processes. Organisations that use OT can be found in many sectors, including water and energy, and often operate a nation's critical infrastructure. These organisations have been under a digitalisation process, which along with increasing regulatory pressures have necessitated changes in their cybersecurity practices. The lack of internal resources has often compelled these organisations to turn to external consultancy to enhance their security. Given the differences between OT and Information Technology (IT) security practices and that OT cybersecurity is still in its infancy, developing a security culture in OT environments remains a challenge, with little research investigating this topic.

We have conducted 33 interviews with professionals with a security related role working in various OT sectors in the UK, on the subject of security culture development. Our analysis indicates three key organisational barriers to the development of a security culture: governance structures, lack of communication between functions, and the lack of OT cybersecurity expertise. Subsequently, the role of consultants and security solution vendors in overcoming these barriers through consultancy is demonstrated. We therefore argue that these stakeholders play a crucial part in the development of security culture in OT and conclude with recommendations for these organisations.

1. Introduction

Organisations that use Operational Technology (OT) have embarked on a digital transformation over the past years, a process known as Industry 4.0 [1], IT/OT convergence [2], or the Industrial Internet of Things (IIoT) [3]. This digitalisation provides many benefits, including reduced costs, and more efficient and accurate data collection [1]. However, it has also increased OT systems' security risks, as OT and IT are becoming more interconnected [2]. As these organisations are often responsible for operating a nation's critical infrastructure, like those in the energy, transport, and water sectors, their cybersecurity is of paramount concern [4].

Operational Technology (OT) refers to systems that control and monitor industrial processes and equipment [5]. Various other terms are used to describe operational technology, including Supervisory Control and Data Acquisition (SCADA) systems and Industrial Control Systems (ICS) [2], with OT being the one most commonly used. Given OT's cyber-physical nature, a cyber-attack can have financial as well as physical impact, leading to injury, loss of life, and environmental damage [6]. Previous such attacks include Stuxnet, which targeted Iran's nuclear capabilities, and the Ukrainian energy system attacks in 2015-16, which resulted in wide-spread power outages [7]. More recently, when ransomware hit their enterprise estate, Colonial Pipeline had to proactively halt their operations over fears that it would spread to their OT estates, which led to fuel shortages [8].

Aside from the increased rates of cyber-attacks on OT, regulation was another factor that has practically forced these OT organisations to enhance their cybersecurity practices. Namely, the EU's Network and Information Systems (NIS) directive, which was passed into United Kingdom (UK) law in 2018, designated organisations operating critical infrastructure as operators of essential services (OES) [9]. Similar measures have been taken in sectors where the NIS does not apply, such as the OG-86 directive for major hazard industries like oil and gas [10], and the International Maritime Organisation's guidelines on maritime cyber risk management [11].

Against this backdrop, attempts to improve the cybersecurity of organisations using OT have necessitated changes in their technology, processes, and people. Nevertheless, OT cybersecurity has followed a similar trajectory to information security [12] with research in OT cybersecurity technologies (e.g., [13]), and accordingly processes (e.g., [14]), reaching a level of maturity that people-related security research in OT has not reached yet [15]. People in OT organisations are often targeted as an initial access vector via techniques like spear-phishing, as is the case in most recorded OT attacks in since 2013 [7]. As such, developing a security culture in OT has been promoted by various in-

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.

USENIX Symposium on Usable Privacy and Security (SOUPS) 2022. August 7 -- 9, 2022, Boston, MA, Canada.

dustrial [16] and governmental bodies [17], since a strong organisational security culture ensures that security is an intrinsic part of employees' duties, and that security is perceived positively and pursued in all levels of management [18].

Most research in security culture has been conducted in IT organisations. Nevertheless, organisations that use OT differ from the ones using IT on their structure, values, as well as technology used. Firstly, these organisations use OT in their various industrial sites, as well as IT in the enterprise part of their business [19]. OT has a lifespan of decades, thus necessitating tailored security practices compared to IT [6]. Likewise, the safety and uptime of their services is of paramount importance, with security only recently becoming a concern [6]. Moreover, various stakeholders have a vested interest in OT cybersecurity, ranging from the government to their supply chain [20]. Additionally, security consultancies and security product vendors are also heavily involved in OT cybersecurity [21].

We therefore follow the argument that different types of organisations need tailored approaches to develop their cybersecurity culture to conduct our research [22]. Organisations using OT are additionally an ideal case study at security culture development as they usually are at early stages of this process. We have conducted 33 interviews with professionals with a security related role in various sectors in the UK, including water, transport, and energy. More specifically, we aim to answer the following research questions in this work:

1. What are the biggest organisational barriers to developing a cybersecurity culture for OT environments?
2. How do security consultants and security solution vendors contribute to overcome these barriers?

Our results demonstrate:

- Three key organisational obstacles towards a security culture: (i) governance structures, (ii) lack of communication between functions, and (iii) lack of OT cybersecurity expertise.
- The role of security consultants and solution vendors in overcoming these obstacles through consultancy, and in turn, influencing the security culture development of these organisations.

Our findings present insights to the research in OT cybersecurity by providing practical recommendations on how common organisational obstacles can be overcome. Additionally, our research contributes to the wider security culture literature by describing the complexities OT organisations face in their attempts at developing a security culture and, more importantly, the role of external stakeholders in shaping this culture.

2. Related work

2.1 Differences between IT and OT

Many significant differences between OT and IT exist, which necessitate tailored security approaches and ultimately affect an organisation's security culture. For example, OT's lifespan, which is typically decades long, complicates its security. Updates for a system might not be available, because the manufacturer might have stopped supporting the product, or in some cases, has ceased operating. Generally, patching and updating practices cannot be directly translated from IT to OT environments, as they must be in continuous operation [6]. This requires patches to be applied in tightly planned maintenance windows which take place a few times a year. Even measures such as longer passwords are not acceptable in time-critical scenarios, where availability and safety concerns are of greatest priority [23].

Aside from the technical differences, organisations using OT differ structurally from IT ones. They have a hierarchical structure, where the enterprise part of the business is separated from the industrial one, both physically and digitally. This separation is demonstrated by the Purdue model, a typical architecture reference model for OT, which consists of three zones and six levels: an enterprise zone and an industrial zone, separated by a demilitarized zone (DMZ) (see Appendix A for a diagram) [24]. Accordingly, different functions with divergent priorities are responsible for the technology and budget of each zone. For example, established information security principles in IT like the confidentiality, integrity, and availability (CIA) triad need to be reshaped to fit OT priorities, by including values such as safety and resilience [25]. Finally, security expertise in OT is relatively scarce. While incidents like Stuxnet have alarmed some organisations on the importance of cybersecurity, it was not until the NIS regulations that most organisations were propelled to act on their OT cybersecurity.

2.2 Security culture background

Security culture is a subculture of the wider organisational culture (i.e., 'The way things are done here' in an organisation). Many culture theories have been proposed, with Schein's model being the most common in security culture research [26]. Accordingly, culture is broken-down into three increasingly observable layers: tacit assumptions, i.e., the values taken for granted in a company, the level of espoused values, and the artefacts and creations level [27]. In the case of security, the assumptions level includes core operational values, which are often taken for granted, such as an organisation's risk-taking appetite. The espoused values level encompasses employees' security attitudes and perceptions. Finally, the observable artefacts level includes

objects like training material and policies and procedures around security [26].

Research has predominantly focused on information security culture with research in cybersecurity culture recently becoming more prominent [22], as cybersecurity encompasses the protection of other assets aside from information, including the people that operate in cyberspace [28]. Nevertheless, as OT security has a strong cyber-physical element, we suggest that security culture is a more suitable and encompassing term for this area of research.

As culture is a construct, a variety of definitions on its constituent elements exist [29] , with the overwhelming majority focusing on attributes such as perceptions, values, attitudes, and behaviours around security [30]. ENISA's definition is a typical example, with culture defined as: 'The knowledge, beliefs, perceptions, attitudes, assumptions, norms and values of people regarding cybersecurity and how they manifest in people's behaviour with information technologies.' [18]. Accordingly, several factors that affect security culture have been proposed in the literature [22]. These include management support, i.e., the involvement and leadership displayed by a company's management, and security policies and their attributes such as accessibility and clarity. More recently, the effect of national culture and regulation have also started receiving attention as culture influencing factors [22].

Nevertheless, the literature on security culture has a few gaps. Firstly, the most prominent culture influencing factors are associated with internal processes of an organisation, with little consideration given to exogenous factors. Aside from the role of the senior management and the security function, the role of different stakeholders in shaping an organisational security culture is often unexplored, especially those outside an organisation such as regulators, governments, or consultancies. Finally, the research area is dominated by theoretical frameworks, followed by quantitative approaches, with qualitative research lagging behind [22]. Our research attempts to fill some of these gaps, by demonstrating the role of other external stakeholders in shaping these companies' security culture, as is the case with consultancies and solution vendors in the OT space. Additionally, qualitative research can provide more in-depth insights in the security professionals' espoused security values, compared to quantitative approaches.

2.3 Security and safety cultures in OT

Safety culture is another part of the wider organisational culture in OT organisations. It became prominent following the nuclear accidents at Three Mile Island and Chernobyl, with the report on Chernobyl's aftermath often cited as the first use of the term [31]. What initially started with a focus on nuclear facilities, has in recent decades spread over into other industrial sectors that use OT including energy, oil and gas, and water. Compared to security, safety is now an established culture in these companies, and its effects are visible in all organisational levels, from small proactive acts like holding the handrail, to safety being a core organisational value, appearing in annual reports and in boards' communications to employees [25]. Nevertheless, there are many commonalities between the two cultures, with factors such as top management support and training considered as important in their development [32].

Research in organisational factors and security culture is an emerging area, with a few researchers looking into the topic in the past decade. Security workers in OT environments were described as 'shadow workers' due to the many obstructions they faced when attempting to fulfil their responsibilities. One such obstacle was that security was perceived as a concern to be exclusively handled by security personnel. Additionally, organisational divisions obstructed the visibility of their function and hindered security communications, further complicating their tasks [33]. Nævestad et al. have assessed the security culture of a critical infrastructure company in Norway. Their second study [34], two years after the first [35], demonstrated that the organisation's security culture had improved, with the authors attributing it to measures such as improved security communications between supervisors and employees.

Dewey et al., in their case studies of four UK nuclear organisations, highlighted various restructuring efforts aiming to integrate security into existing business structures. Additionally they demonstrated various challenges faced by security employees in their efforts to improve their organisation's security culture [31]. For example, mediums such as email were not as effective in distributing security communications, due to the non-office nature of many employees (e.g., rail operators, engineers, maritime). Finally, given the prevalence of safety culture, employees were more appreciative of the need for safety compared to security.

The impact of NIS in several UK sectors, including water [36] and energy [37] has also been investigated. Given its infancy, considerable inter-organisational collaboration was undertaken to transpose NIS into sectoral contexts, through various self-organising networks [20]. For example, the NIS necessitated closer collaboration between competent authorities and governmental entities, to translate the NCSC's Cyber Assessment Framework (CAF) to the needs of each sector. Other instances of inter-organisational collaboration include working groups between OT companies and critical suppliers where common security requirements were examined. Finally, Michalec et al. [36], in their case study of the water industry, have proposed that these collaborations were also influential in shaping the wider water sector's governance strategies.

These collaborations helped improve the sectoral understanding of the NIS, contributing to the knowledge and understanding of OT cybersecurity issues, and in turn,

strengthened both the sectoral and national security cultures in the UK. However, there is still little work on how OT cybersecurity knowledge and understanding are developed in an organisational context. As evidenced by these intra-organisational collaborations, cybersecurity in OT companies depends on various stakeholders, including the government, competent authorities, original equipment suppliers, and system integrators. Additionally, our research aims to highlight the role other external stakeholders have in shaping this culture, namely, security consultants and security solution vendors.

2.4 Consultancy background

Organisations lacking expertise in a topic [38] or facing a lack of professionals in the employment marketplace often turn to consultancy to fill these gaps [39]. Knowledge sharing, the delegation of functional activities, and the design of processes and procedures are among some of the potential outcomes of the consultancy process [40]. Consultancies vary in sizes and offerings, including mega consultancies offering a variety of services (e.g., tax, accounting, IT etc.), independent consultancies specialising in fewer areas, and vendor consultants whose services relate to the support of their software and hardware offerings [40]. Generally, consultants have been described as 'therapists', 'doctors', and 'gurus' in the literature, and are often seen as 'obligatory passage points' supplying expertise to organisations [41]. OT cybersecurity is currently one such area where consultancy is recognised as the first step towards cybersecurity maturity [21].

Consultancy in IT, often focused on the implementation of Information Sharing (IS) [42] and Enterprise Resource Planning (ERP) systems [43], is one area close to cybersecurity which has received considerable academic attention. Broadly, the extant literature on consultancy can be summarised under the following topics: factors that contribute to consultancy's success [44], the client-consultant relationship [41], and the consultants' roles which can range from change agents to uncertainty managers and fashion setters [45], [46]. In the case of consultants as change agents, their role in knowledge sharing has often been recognised, with research demonstrating that it is more probable that an organisation's personnel will value knowledge from external sources compared to internal ones [47], especially when an organisations' own internal capabilities are lacking [39].

Cybersecurity research into consultancy compared to IT is scarce. While the cybersecurity and IT implementation consultancy processes have many commonalities, the literature in IT implementation often emphasizes the value of IT transformation in terms of cost reduction, increased effectiveness etc. [46]. The value of cybersecurity on the other hand is not as clear-cut, with companies regarding cybersecurity as an additional cost. Nevertheless, consultants have been described as cyber advocates i.e., individuals who can persuade organisations to adopt positive security practices [48]. Empirically, Gale et al. have found consultancy to be a driver influencing companies' cybersecurity decisions [49]. Finally, Poller et al. have investigated the effect of external consultancy on the organisational practices of a company's software development groups [50]. While the consultancy had some positive short-term effects, an overall lack of long-term sustainable change was reported.

3. Methodology

We have conducted a qualitative study through semi-structured interviews with professionals with a security-related role from a variety of OT sectors. The research has been approved by the authors' institutional ethics committee. The following sections provide more details on the sampling and recruitment process, interview conducting, and data analysis.

3.1 Sample and recruitment

We have employed theoretical sampling, with participants identified based on gaps in the collected data, or to explore emerging concepts [51]. Additionally, we employed snowball sampling by asking participants to refer us to other potential participants based on their role, which resulted in 11 of the 33 interviews [51]. Participant recruitment was undertaken via LinkedIn. We have decided to stop conducting interviews once our findings have reached theoretical saturation [51], i.e., interviews did not provide any new categories of inquiry or relevant data with respect to organisational barriers and consultancy's role in overcoming them.

While the water sector was our initial focus, the first few interviews led to the choice of including additional sectors, as our participants perceived that most OT sectors were facing similar cultural and security challenges. Our interest into external stakeholders also arose after the first few interviews, as the heavy presence of consultants and security solution vendors in the OT cybersecurity space was made apparent. Finally, our sample choice has presented a few obstacles. The population size is relatively small as organisations that use OT are bounded by regulation and geography, as is the case with UK water organisations which are effectively regional monopolies. Additionally, given the critical nature of operations of these organisations, some secrecy and hesitation were expected. Accordingly, research topics were tailored to be as non-intrusive, based around organizational structures, personnel's attitudes and perceptions etc., as to not be perceived as overtly sensitive by the study's participants.

3.2 Interview design and data collection

Interviews were conducted via Microsoft Teams between July 2022 and January 2023, lasting on average around an hour, and ranging from 45 to 70 minutes. Participants were not compensated for their contribution. Two participants opted to not be recorded, and therefore, data were collected through note taking. The participants' pool includes professionals working in OT companies with roles such as Chief Information Security Officers (CISOs), security managers, and OT managers, as well as external stakeholders including consultants and regulators. With respect to consultancies, we have included participants from large consultancies, smaller consultancies focused primarily on OT cybersecurity, and vendor ones. Overall, our study includes the views of 33 participants from 25 different organisations. A full breakdown of the participants' role and sector can be found in Appendix B.

Semi-structured interviews were used as their flexibility allows the point of view of the interviewee to come across more predominantly than other interview methods such as structured ones [51]. Often, questions did not follow the guide's outline and going on tangents was encouraged, as it provided an indication of what participants think is relevant. Interview guides were tailored based on the participant's role and a sample guide can be found in Appendix C.

3.3 Data Analysis

Microsoft Teams' built-in recording tool automatically produces a transcription. As such, we have used that tool to reduce the time needed to transcribe interviews, compared to a transcription done fully by ear. To code and analyse the transcriptions, the NVivo 12 software was used.

Thematic analysis was chosen to analyse the data, following the recommendations by Braun and Clarke [52], who provide a six-step process, and Ryan and Bernand [53], who provide techniques on how to identify themes. The first author coded all the data, following an open-axial-selective coding process [54]. This required an initial familiarisation with the data, which was aided by theoretical memoing. An open, primarily inductive process was used to develop an initial codebook. Accordingly, through an iterative process, codes were discussed with the other authors in weekly meetings, leading to further refinement, and eventually, to the themes presented in this work.

Using thematic analysis allowed us to identify themes both deductively, i.e., in-line with the literature, including factors such as policies and procedures, but primarily through induction, i.e., from an analysis of the data, such as the differences in mindsets and values between engineers and IT/security personnel. Examples of relevant codes can be found in Appendix D.

3.4 Limitations

We had to resort to online means for recruiting participants, as data collection was conducted during the Covid pandemic. In some cases, our attempts were perceived as social engineering, which was potentially amplified by campaigns such as 'Think before you link', by the National Protective Security Authority (NPSA) in the UK [55]. Paired with the overall small population size, a considerable amount of time and effort was spent to reach potential participants and obtain their trust.

Moreover, while our data have been collected from a variety of security professionals both internal and external to these organisations, and from various OT sectors in the UK, we do not claim a high level of external validity. Other nations, with different regulations on the cybersecurity of OT, or where organisations using OT have different operational models (e.g., utilities are publicly owned), might not necessarily face the same organisational obstacles, or the role of consultancy might be mediated by other stakeholders such as the government.

4. Results

4.1 The OT security landscape

The introduction of the NIS regulations and other similar directives prompted organisations using OT to actively take measures to improve their cybersecurity, leading to the realisation that becoming cyber-secure would be a challenging task. Many of these organisations have been reaping the benefits of digitalisation for years, without securing their technology, further complicating this task. As such, organisations often had a *"knee-jerk" (P29)* reaction, allocating OT cybersecurity responsibilities to their IT function, or their security function which would still be IT-focused. Given the lack of resources and inability to cultivate OT cybersecurity expertise internally, as well as the urgency of the matter due to the newly introduced cybersecurity directives, external consultancy was often the solution. Consultancies recognized this business opportunity - *"their eyes light up at the mention of OT security" (P09)* - and moved swiftly to fill this gap. Security solution vendors also found themselves in a similar position of business opportunity to provide consultancy services.

The rush to secure OT systems, combined with the lack of OT security expertise, allowed poor quality security practices to proliferate. Often a company supplying OT cybersecurity services would have never *"stepped foot" (P27)* below the DMZ, both physically, by visiting OT sites to understand their equipment and processes, as well as digitally. Substandard technical solutions, such as poorly designed network infrastructures, weak segregation between IT and OT networks, and penetration testing were often mentioned. Namely, penetration testing would often be a simple vulner-

ability assessment, without any contextual information on how that affects operations. Moreover, it would rarely go beyond the DMZ, limited to testing the segregation from the enterprise to the manufacturing zone. While penetration tests in OT carry substantial risks given the concerns about availability and the presence of legacy equipment, their limited scope provided OT customers with a false sense of assurance.

This lack of understanding extends beyond technology to a cultural level. OT personnel have been trained for years to value the availability and safety of their systems, where Service Level Agreements (SLAs) exist on the amount of allowed downtime. Additionally, safety is of paramount importance, with many systems being safety critical. On the other hand, IT and security professionals are eager to apply changes to secure OT systems, often by trying to directly translate from the IT to the OT world. Ultimately, the insufficient experience in OT environments has led to inferior quality risk assessments, where the cyber-physical nature of these systems was often lost, and an attack's impact on safety and the environment was not being considered:

"The experts that are brought in to do some of the risk assessments are maybe 27001 [standard] *qualified and they look at it* [rolling stock] *almost as an IT system. While there's a lot of overlap, 62443 incorporates the safety side as well and includes that with the CIA risk factors. And that's something that quite often we have to go back and say, well, you haven't considered safety on this one. What's the knock-on effect?" - P30, Consultant*

Note – ISO/IEC 27001 is an international information security management standard, whereas IEC 62243 is the equivalent series of standards for OT cybersecurity.

4.2 Organisational barriers

While the overall security maturity in OT organisations is improving, we have identified three recurring organisational obstacles that pose a challenge to the development of a security culture in OT: (i) governance structures, (ii) lack of communication between different functions, and (iii) lack of OT cybersecurity expertise.

4.2.1 Governance

Cybersecurity in OT is constrained by the operational models of these organisations. Typically, the operations and engineering functions are responsible for the industrial sites and their operational technology. IT and security sit on the enterprise side of the business, having historically being tasked with its information security. As these functions have grown organically over the years, and cybersecurity for OT has only recently become a concern, this governance model complicates security knowledge exchange and communication efforts.

These functions typically have different reporting lines, leading to a situation where managers do not speak directly with their counterparts but use a chain of commands where information must travel upwards and then downwards. This often allows other organisational politics to get involved, and information can get diluted. Additionally, the number of people involved in a cybersecurity decision, including personnel from both IT and OT responsible functions, also hinders the decision-making process.

Another obstacle is the ownership of operational assets. While the IT department is usually tasked with a company's cybersecurity, the budget and resources for OT are typically owned by the operations function, which has different priorities on how they should be spent. Participant 02 recalled the pushback they had received when a decision to install an intrusion detection solution their OT estate was made:

"That got a lot of pushback with the OT teams because it's quite expensive. When you look at some of the IT technologies, spending half a million pound on something is neither here nor there, that's general. But when you spend half a million pound in the OT is 'Why are we spending so much on that? That's ridiculous. That's technology. I could spend that and repair all these different systems and repair this and repair that'." – P02, OT Manager

Finally, the problem of ownership and governance is intensified in sectors which operate offshore and onshore assets, as they can be subject to different cybersecurity regulations or accountable to different regulatory authorities. Similarly, in sectors like oil and gas, joint ventures and the outsourcing of the operations are common practices, which often leads to certain stakeholders having a disproportionate amount of responsibility. The international operations of some organisations also complicate cybersecurity, as is the case with the maritime industry. While cybersecurity risk assessments were made mandatory by the IMO since 2021, there is no guarantee that this is upheld by the relevant national authorities - *"There's not many countries taking that seriously". - P28, Academia and industry coordinator*

4.2.2 Lack of communication

The lack of communication between functions with cybersecurity responsibilities was another commonly referred barrier obstructing the development of a security culture. The non-desk nature of many OT roles, along with practices like shared workstations and user accounts limit the effectiveness of communication mediums such as e-mails or intranets in delivering cybersecurity relevant information. This is further exacerbated by the existing governance structures, as well as the lack of OT experience. As P15 bluntly stated:

"If you talk to the IT department, they don't have any expertise in OT and they don't really talk to the engineering de-

partment because the engineering department doesn't want to talk to IT people." - P15, Security solutions vendor

The different value systems or *"mindsets"* of these two groups also contribute to this communication barrier. The *"engineering mindset"* prioritizes the stability and service of their systems. Additionally, both occupational and functional safety are paramount given the potential physical consequences of safety incident. Accordingly, engineers are *"conservative with a little c" (P29)* when it comes to changes in their equipment and practices. Often, a reticence by OT engineers exists in installing new equipment in OT systems given the potential disruption they can cause – *"Something is working. Don't touch it" (P29).* On the other hand, IT and security professionals with a *"technology mindset"* are typically more familiar with newer technologies and more relaxed on their implementation. As such, IT and security experts are often perceived as intruders in OT environments, with both teams viewing each other as hindrance.

The different mindsets and lack of communication can lead to futile attempts at securing OT, with the example of OT professionals not allowing changes in their environments being regularly mentioned. Moreover, substandard security practices have damaged the trust between these two functions and diminished the willingness for future collaboration. Participant 22 recalled their experience with unimplementable directives coming from the IT department in a previous engineering role they had:

"We got directives through about centralised antivirus update for your OT systems, password protection policies and things like that. And you went trying to implement, things like IT security experts [saying] *this is what your password complexity is. Read up on it, sent an email 'Sorry but Windows can't do that'.*

It's as silly as that, and these are IT security experts and they're asking you to do something that Windows can't even do. It's an example of things that filtered down. And it starts making you go; 'Am I going to read and do everything they send down to me now? Probably not'." - P22, Consultant

4.2.3 Lack of expertise

The third major barrier towards an OT security culture is the lack of OT cybersecurity expertise, with both governance and communications issues obstructing its development. In turn, the lack of expertise leads to communication obstacles and diminished trust between stakeholders with cybersecurity responsibilities. By defining expertise as a measure of knowledge and experience, it follows that expertise is not easily achievable unless there is sufficient hands-on experience in a specific context [56]. In the case of OT cybersecurity, this would entail both cybersecurity experience, but more importantly, experience in industrial environments using OT.

Nevertheless, there are a few challenges that prevent this expertise gap from closing, aside from the field's current immaturity. One such challenge is the different professional routes towards a job in engineering compared to cybersecurity. Engineers and other OT professionals *"have come up on the tools" (P27),* primarily through vocational education such an apprenticeship after finishing secondary education. OT personnel may hold an engineering-related degree or diploma, but a master's degree is rarely a prerequisite for such roles. On the other hand, security is becoming a profession where a degree is increasingly desired, compared to previous decades where information security had been creeping into the role descriptions of IT professionals. In our sample, at least five participants had obtained a security or technology-related degree later in their career, which facilitated their move to a cybersecurity role.

In the case of OT, cybersecurity was typically added to personnel's existing job descriptions, without these organisations firstly supplying resources to train their personnel on cybersecurity. OT personnel who for years have been primed to value the safety and uptime of their systems suddenly also had to value cybersecurity. Accordingly, the need for cybersecurity would not be appreciated, especially if there were no targeted efforts to communicate its value and link it to their priorities. Moreover, incorporating security in their everyday tasks adds to their workload, and as such, they will often pushback to such changes. Similarly, the IT and information security side was often additionally tasked with the cybersecurity responsibilities for OT. Realising their lack of expertise and the difficulty in setting up relationships and communication corridors with the operations and engineering functions, accepting these responsibilities is a difficult choice:

"CISO's five years ago did not have oversight on the onboard stuff, and now it's come into question there is a kind of defensive, you know, how do you admit you were wrong in the past? Also, if you're a CISO with limited budget and pressures on your existing infrastructure, to willingly lift the lid and say, 'All this new stuff, [is] *now in my scope and I'd have to argue for more budget', that just makes your life more difficult as well." – P26, Security solutions vendor*

Finally, another factor which complicates the closing of this expertise gap is the preference companies have on hiring professionals from their sector. Participant 15 recalled an interaction with an oil and gas company:

"And then when they were bemoaning, you know, the lack of skills and the lack of people who they could get involved in cyber security. 'We can do that, we've got some great guys working in shipping now who would be really good for your

offshore installations'. And the answer I got was: 'So not oil and gas men then'." – *P15, Security solutions vendor*

4.3 The role of external experts

This section discusses the role of security consultancies and solution vendors in overcoming these organisational barriers. Consultancies and vendors operate across a variety of sectors and their accumulated experience on the threats and security solutions for OT is sought after by OT organisations. Additionally, their expertise in both the engineering and cybersecurity domains makes them efficient mediators between organisational functions. As these stakeholders commonly work with OT organisations at the initial stages of their cybersecurity journey, we argue that they shape OT organisations' thinking and actions around cybersecurity, and therefore, play a part in shaping their security culture. Security consultancy can take many forms, varying from long-term contracts where consultants are embedded into the client organisation, to shorter-term contracts with a specific focus (e.g., auditing, regulation). The unique challenges faced by OT organisations such as the disperse nature of their assets and legacy equipment, and the fact that IT security practices cannot be directly translated to OT have created a market for OT-specific solutions. However, our study's participants working in the security solutions industry acknowledged that the need for OT security solutions is not always appreciated by the market.

"Why aren't we selling more? Because the market doesn't understand so we have to get the market to understand so that we can sell more." – *P15, Security solutions vendor*

The theme of *"educational sales"* was a common occurrence in our interviews, where potential customers need to be taken through a journey of understanding of their underlying needs before a sale can take place and solutions are implemented and supported. As security solutions are not a one-off purchase, security vendors often resort to consultancy to improve their potential customers' understanding of security.

4.3.1 Overcoming organisational barriers

Making sense of the governance and responsibility structure of their customers is typically the first task these stakeholders undergo, as OT security responsibilities are not always clear cut:

"The first thing you have to work out is who's doing what and who does the company think is running their OT security." - *P15, Security solutions vendor*

The value of a proposed solution then needs to be demonstrated to the relevant stakeholders. However, given the lack of communication and understanding between different functions, these external stakeholders are often asked to support the IT and/or operations teams to present a solid business case to their management. Moreover, given the internal lack of expertise, external experts have a detrimental role is in effectively delivering these solutions, be it technical or procedural. Additionally, these experts can influence changes in governance, by guiding new teams such as joint IT and operations functions. They also guide the allocation of cybersecurity responsibilities; both internally, by advising on the responsibilities for different roles, as well as intra-organisationally, in the case of joint ventures between multiple companies.

Consultants and vendors also act as translators between different teams. Having observed the confusion caused by different uses of common terminology, such as TTL which means time-to-live, transistor-to-transistor logic, or threat-to-life to different stakeholders, one participant had created a glossary to improve the understanding of the security-responsible functions. More broadly, these external experts often sit in the middle of different teams, acting as facilitators and helping build bridges between them. Participant 18 recalled their experience during a workshop where communication problems between functions were present:

"It was very clear from day one … that there was this issue because it was all engineering and operations on the left of the room and IT was on the right and they were always just sort of looking over and then sort of switching their heads back… It was obvious that there was some clashes and some politics there. … So that's when the consultants would come in and do a bit more facilitation and say, well, have you thought about this … and you'd play the sort of the mediator between them. And more often than not, it actually becomes resolved."- P18, Consultant

Finally, consultants and security vendors contribute to the security culture of an organisation by equipping employees with an understanding and knowledge about OT cybersecurity. There is a multitude of ways for knowledge exchange to happen, as these experts are typically brought in an organisation with low cybersecurity maturity. Aside from bringing various teams together and increasing their cohesion, they advise on suitable technical security solutions. Moreover, they work on developing policies, as is the case of equipment procurement, pushing cybersecurity requirements into the supply chain. In other cases, they collaborate with communication teams to distribute security awareness material, or mentor OT professionals towards a cybersecurity certification or degree.

Given the continuously evolving nature of cybersecurity, both in terms of technology and threat landscape, cybersecurity services should be supported and reviewed on a continuous basis. As these collaborations are usually long-term, the concept of taking customers on a journey was often referenced by participants, necessitating the development of long-term relationships between these stakeholders and their

customers. Several factors can affect their success, with security professionals with an engineering background feeling that they could more easily appreciate OT engineers' needs and earn their trust. This also extends to the organisational level, with consultancies with engineering expertise finding it easier to build relationships with OT staff, given their common backgrounds.

"I'm probably a bit better at it because I've come from that background, so I kind of understand their way they do risk assessments". – P22, Consultant

Nevertheless, participants often recognised that their attempts at building these relationships will be futile if relevant stakeholders have not been engaged by their organisation beforehand. Additionally, they can be met with distrust from other functions, or individuals, as they are perceived as part of the team that has contracted them. Participant 12 recalled their experience with a client organisation:

"There was big hostility for what we were trying to do to the point we were doing assessment questionnaires and they [the industrial site] *were being deliberately evasive.... For whatever reason, there was a huge distrust between the asset and headquarters." – P12, Consultant*

Finally, technical solutions themselves contribute to the security culture of these companies. Whether it is an asset discovery tool, a network gateway, or an intrusion detection system, technical solutions give OT organisations *"a window into their networks that they didn't have before" (P15),* by exposing new, security related information. For example, asset discovery is a substantial challenge for OT companies, as assets have been accumulating over time and are dispersed in vast geographical areas. As such, asset discovery tools are the first step towards understanding the presence of a variety of operational equipment in OT estates. Similarly, network monitoring solutions can produce alerts on the state of OT networks, allow only specific communications via whitelisting, or alert operators about anomalous behaviours. This information can then aid the security and operations teams to make more educated decisions on how to prioritise and distribute their budgets, ultimately enabling better quality security risk assessments.

5. Discussion

We have identified a few restructuring efforts in these OT organisations through our study, including a merger of operations and IT and the creation of an independent cybersecurity function, all aiming to improve the organisational management of cybersecurity. However, restructuring is not an easy task, especially when departments have grown organically over time. Targeting the lack of communications and expertise can improve an organisation's security culture without a disruptive and costly restructuring effort. Initiatives like workshops and steering groups are a common practice, enabling cross-pollination between various stakeholders. Nevertheless, care should be taken when deploying such initiatives, as it is crucial that the divergence in values and terminology between OT and IT personnel are addressed beforehand.

Given the infancy of the field, the lack of OT cybersecurity expertise is a harder challenge to overcome compared to communication issues. Nevertheless, organisations should aim to close this gap by being less reluctant to hiring OT experts from other sectors, as well as investing in training their OT personnel. Due to the increased digitalisation of OT, IT equipment is increasingly used in OT environments, making OT cybersecurity more accessible to outsiders. However, we suggest that expertise in OT is harder to obtain compared one in cybersecurity. This is primarily due to the different "mindsets" between engineers and IT personnel. Appreciating the engineering way of working and their concerns was often cited as a difficult challenge for IT-based professionals. While both approaches work; IT security professionals moving to OT, and OT professionals moving to security, the latter requires less effort. As such, aside from cross-sectoral hiring, organisations can invest in developing their OT personnel's cybersecurity expertise, by sponsoring OT professionals towards a cybersecurity certification or degree, as well as introducing security training at the early stages of engineer's career, in apprenticeships and other vocational schemes.

Previous research in OT organisations has proposed that they were under heightened pressure to acquire such expertise, with OT cybersecurity lacking a typical career trajectory [19]. Our findings suggest that this situation has been improving, with security roles and responsibilities becoming more standardised, and security becoming more prominent as an organisational function. Nevertheless, this expertise gap will continue posing a challenge and will be further amplified given increased government pressure and proposed changes in regulation. The NIS 2 which will replace current regulation, expands the scope of what constitutes an operator of essential services by including organisations from sectors like wastewater, and provides additional powers to competent authorities [57]. Consequently, an increased number of organisations will be looking to acquire expertise from a limited pool of available talent, further amplifying the need for organisations using OT to tackle this challenge by training their existing personnel.

The identified obstacles of governance, expertise, and communications commonly recur in organisational cybersecurity contexts [58], [59]. These obstacles are often intertwined [60], and have been shown to affect individuals' security behaviours [61], as well as an organisation's security culture [22]. For example, our analysis demonstrates that the lack of OT knowledge by IT professionals leads to diminished trust

and ineffective communications between the IT function and their OT counterpart. At the same time, the lack of communication between these functions impedes the sharing of OT and cybersecurity knowledge. Overall, security practitioners can apply many of the culture literature's recommendations more or less directly, such as two-way communications between different stakeholders [62], and the development of non-technical skills (e.g., communication, leadership, etc.) by security professionals [62], [63], targeting the issues of communication, and subsequently, knowledge sharing.

However, the differences between the organisations using OT and IT should not be overlooked, especially when developing a culture of security in their industrial zones. While most research on communications focuses on the exchanges between the security function and end users [64] or senior stakeholders [65], our analysis shows that in OT organisations communication and knowledge exchange between the functions responsible for IT and OT are a prerequisite to ensure optimal cybersecurity practices. Research in knowledge and awareness also focuses on cybersecurity education of various personnel [66], and simultaneously the need for security professionals to understand personnels' priorities and work processes [64]. This dialectic process in turn allows for the effective tailoring of security procedures, to better suit personnel's workload and organisational goals, preventing shadow security practices [64].

Nevertheless, the technology underlying most human-centred cybersecurity research is IT, in which security professionals are experienced. The obstacle of security knowledge is amplified in OT contexts, as security practitioners need to understand the underlying operational technology, as well as its end-users' priorities (e.g., safety, availability). Usability research has demonstrated how OT users' security perceptions are affected by the design constraints of their equipment, and their familiarity with IT equipment [67]. Additionally, our research demonstrates how the different "mindsets" between OT and IT stakeholders act as a barrier towards effective cybersecurity, by undermining the trust and collaboration between these functions, and thus hindering communications and knowledge exchange.

Decisions at the organisational level are also affected by the potential for physical impact caused by an OT incident (e.g., loss of service, injury). As such, OT-centred organisations have developed a culture of safety over the years [68], which is uncommon in IT-based organisations. The OT digitalization has also brought the prospect of a cyber incident causing physical damage. However, the relationship between the two cultures is still unclear. Future research could investigate the extent to which these two cultures overlap, or how the predominant safety culture (e.g., perceptions, attitudes) affect the security culture both at an organisational and managerial level, as well as at the individual level.

According to our analysis, external stakeholders impact the security culture of OT organisations, with their primary contribution being knowledge transfer at a point where most organisations have low OT security maturity. Previous research has demonstrated external stakeholders' impact on a company's knowledge and organisational practices [69], with knowledge communicated through various informal and formal mediums (e.g., steering groups, conversations, training, policies and procedures) [40], as also demonstrated through our analysis. To our knowledge, our study is one of very few looking at the effect of consultancy on cybersecurity practices, and accordingly, security culture in organisations. Generally, the consultancy processes described in this work have led to changes in how cybersecurity is perceived and managed in OT organisations. This is partly owned to the elevation of cybersecurity to a visible business goal in recent years, which has strengthened the remit of change for these external stakeholders. This contrasts with Poller et al.'s case study [50], where it was recognised that the consultants' remit did not explicitly include advising on organisational practices, thus failing to have long term impact.

Aside from their positive contributions, we have also observed a move of OT cybersecurity experts to consultancies during our research. The acquisition of talent by consultancies which can offer higher salaries and other benefits is hurting organisations that use OT, as it further limits the pool of OT expertise they can tap into. Moreover, substandard security works by consultancies and solution vendors were also commonly referenced in our participants' responses. Given the limited security understanding of many of these organisations, work from these external stakeholders can provide them with a false sense of assurance. Accordingly, this leads to a situation where organisational security culture cannot flourish, as cybersecurity is perceived as an issue that was addressed by external stakeholders.

Organisations can partially prevent these substandard solutions by building their 'intelligent customer' capability [70], i.e., obtaining an adequate level of security knowledge and a wider understanding of how security fits into their operations. This in turn enables organisations to make informed choices when outsourcing their security, as well as being able assess the quality of the work delivered. This is a sensible approach for many aspects of security, including the procurement of services such as intrusion detection systems (IDS) or security operation centres (SOC). However, while a company's technology and processes are often tailored by external professionals, and security knowledge is transferred to the relevant functions, other, softer sides of culture, need to be developed in-house.

Namely, the role of management and security communications are two essential culture-affecting factors which cannot be as easily influenced by these external stakeholders [71]. The top management's role in developing a security culture is commonly referenced in the literature [22], with our participants also agreeing that this top-down approach is necessary, especially at the current stage where organisations have only recently started improving their cybersecurity practices. Interventions, coordination, and communication from the upper echelons of a company must be present to engage employees and convince them about the importance of cybersecurity [72]. The role of direct supervision is also important, as employees' security perceptions are impacted by the prioritization of security by their direct managers [73]. Finally, existing communication channels and methods need to be leveraged with language that is familiar to employees, to embed security into other core organisational values including safety and the provision of essential services such as clean water or electricity.

All in all, external security experts have the potential to greatly benefit organisations using OT. They set strong foundations through designing security processes and procedures and improving their technology. Additionally, they are influential in shaping various factors regarded as important to enhance a security culture, including liaising with an organisation's management to coordinate security efforts. More importantly, their collaboration with cybersecurity responsible personnel can directly affect their personnel's attitudes and perceptions around security.

Nevertheless, organisations need to have the absorptive capacity to exploit this external knowledge [74], by obtaining it, and assimilating it internally [75]. We have demonstrated how external experts through the process of consultancy are crucial in this knowledge exchange with organisations using OT. However, as cybersecurity is not a one-off purchase, organisations using OT should make active efforts to communicate the need for security by considering the different mindsets and values of OT employees, as well as by providing relevant security awareness and training, to assimilate this knowledge. This in turn, can lead to an enhanced security culture, where security becomes embedded into everyday processes and practices, as well as employees' duties.

6. Conclusion and recommendations

Cybersecurity for OT is a fast-growing area, requiring organisations using OT to make drastic changes of their practices, technologies, and people. Accordingly, these changes constitute the first step towards developing a security culture. Through our analysis of 33 interviews with professionals with a security related role in the OT space, we have identified three key organisational obstacles: governance, lack of communication, and lack of expertise. Moreover, we have demonstrated the role of security consultants and security solution vendors have in overcoming them. Consequently, these stakeholders set some of the foundations for developing this culture by breaking down communication and knowledge barriers and shaping various culture affecting factors such as policies and procedures. Overall, our work highlights the role external stakeholders have in the development of an organisational security culture, an aspect that is overlooked in the security culture literature.

While these external experts contribute to the early stages of culture development, organisations using OT need to be able to absorb their expertise and expand the scope of their efforts to achieve a strong security culture. Future research could investigate which conditions make an organisation better at absorbing this external knowledge, and how its assimilation can lead to a stronger security culture. As such, we conclude with three recommendations for organisations using OT.

1. Target different employees based on their roles and "mindsets" on the need for cybersecurity. Accordingly, meditate these differences to allow for improved understanding between functions and increased knowledge absorption both from external sources as well as inter-departmentally.
2. Rather than over relying on IT-based security expertise, training OT personnel in cybersecurity at various stages of their career through apprenticeships, certifications, and degrees, can help accelerate the closing of the expertise gap.
3. The use of external expertise through consultancy and solution vendors can help build strong foundations for cybersecurity, but it takes more to cultivate a security culture. A top-down effort to convey the need for cybersecurity to the various functions responsible and targeted communications are especially important to increase your organisation's absorption capabilities before efforts are made to assimilate this knowledge and enhance your security culture.

Acknowledgments

This project was funded by the UK EPSRC grant EP/S022503/1 that supports the Centre for Doctoral Training in Cybersecurity delivered by UCL's Departments of Computer Science, Security and Crime Science, and Science, Technology, Engineering and Public Policy.

We would also like to thank the four anonymous reviewers whose insightful comments and suggestions have improved this paper.

References

[1] Symantec, 'Smarter Security for Manufacturing in The Industry 4.0 Era', 2017. https://docs.broadcom.com/doc/industry-4.0-en (accessed Aug. 15, 2022).

[2] U. P. D. Ani, H. (Mary) He, and A. Tiwari, 'Review of cybersecurity issues in industrial critical infrastructure: manufacturing in perspective', *Journal of Cyber Security Technology*, vol. 1, no. 1, pp. 32–74, Jan. 2017, doi: 10.1080/23742917.2016.1252211.

[3] H. Boyes, B. Hallaq, J. Cunningham, and T. Watson, 'The industrial internet of things (IIoT): An analysis framework', *Computers in Industry*, vol. 101, pp. 1–12, Oct. 2018, doi: 10.1016/j.compind.2018.04.015.

[4] 'Critical National Infrastructure | CPNI'. https://www.cpni.gov.uk/critical-national-infrastructure-0 (accessed Aug. 01, 2021).

[5] 'Definition of Operational Technology (OT) - Gartner Information Technology Glossary', *Gartner*. https://www.gartner.com/en/information-technology/glossary/operational-technology-ot (accessed Feb. 12, 2023).

[6] N. Tuptuk and S. Hailes, 'Security of smart manufacturing systems', *Journal of Manufacturing Systems*, vol. 47, pp. 93–106, Apr. 2018, doi: 10.1016/j.jmsy.2018.04.007.

[7] T. Miller, A. Staves, S. Maesschalck, M. Sturdee, and B. Green, 'Looking back to look forward: Lessons learnt from cyber-attacks on Industrial Control Systems', *International Journal of Critical Infrastructure Protection*, vol. 35, p. 100464, Dec. 2021, doi: 10.1016/j.ijcip.2021.100464.

[8] Sean Michael Kerner, 'Colonial Pipeline hack explained: Everything you need to know', *WhatIs.com*. https://www.techtarget.com/whatis/feature/Colonial-Pipeline-hack-explained-Everything-you-need-to-know (accessed Feb. 12, 2023).

[9] NCSC, 'NIS introduction', 2019. https://www.ncsc.gov.uk/collection/caf/nis-introduction (accessed Aug. 15, 2022).

[10] 'Cyber security - Electrical, Control and Instrumentation (E, C&I) - HSE'. https://www.hse.gov.uk/eci/cyber-security.htm (accessed Feb. 12, 2023).

[11] 'Maritime cyber risk'. https://www.imo.org/en/OurWork/Security/Pages/Cyber-security.aspx (accessed Feb. 12, 2023).

[12] 'The 5 Waves of Information Security – From Kristian Beckman to the Present | SpringerLink'. https://link.springer.com/chapter/10.1007/978-3-642-15257-3_1 (accessed Feb. 12, 2023).

[13] J. Suaboot *et al.*, 'A Taxonomy of Supervised Learning for IDSs in SCADA Environments', *ACM Comput. Surv.*, vol. 53, no. 2, p. 40:1-40:37, Apr. 2020, doi: 10.1145/3379499.

[14] Q. S. Qassim, N. Jamil, M. Daud, A. Patel, and N. Ja'affar, 'A review of security assessment methodologies in industrial control systems', *ICS*, vol. 27, no. 1, pp. 47–61, Mar. 2019, doi: 10.1108/ICS-04-2018-0048.

[15] S. Evripidou, U. D. Ani, J. D McK. Watson, and S. Hailes, 'Security Culture in Industrial Control Systems Organisations: A Literature Review', in *Human Aspects of Information Security and Assurance*, in IFIP Advances in Information and Communication Technology. Cham: Springer International Publishing, 2022, pp. 133–146. doi: 10.1007/978-3-031-12172-2_11.

[16] DCMS, 'Water Sector Cyber Security Strategy', p. 12.

[17] NCSC, 'A positive security culture'. https://www.ncsc.gov.uk/collection/you-shape-security/a-positive-security-culture (accessed Nov. 27, 2021).

[18] ENISA, 'Cyber Security Culture in organisations'. https://www.enisa.europa.eu/publications/cyber-security-culture-in-organisations (accessed May 31, 2021).

[19] O. A. Michalec, D. van der Linden, S. Milyaeva, and A. Rashid, 'Industry Responses to the European Directive on Security of Network and Information Systems (NIS): Understanding policy implementation practices across critical infrastructures', presented at the Sixteenth Symposium on Usable Privacy and Security (SOUPS 2020), 2020, pp. 301–317. Accessed: Feb. 01, 2022. [Online]. Available: https://www.usenix.org/conference/soups2020/presentation/michalec

[20] T. Wallis and C. Johnson, *Implementing the NIS Directive, driving cybersecurity improvements for Essential Services*. 2020, p. 10. doi: 10.1109/CyberSA49311.2020.9139641.

[21] Idaho National Laboratory, 'Building an Industrial Cybersecurity Workforce, A Manager's Guide'. Idaho National Laboratory. Accessed: Feb. 21, 2023. [Online]. Available: https://inl.gov/wp-content/uploads/2021/02/ICS_Workforce-ManagersGuide2021.pdf

[22] B. Uchendu, J. R. C. Nurse, M. Bada, and S. Furnell, 'Developing a cyber security culture: Current practices and future needs', *Computers & Security*, vol. 109, p. 102387, Oct. 2021, doi: 10.1016/j.cose.2021.102387.

[23] O. Michalec, S. Milyaeva, and A. Rashid, 'When the future meets the past: Can safety and cyber security coexist in modern critical infrastructures?', *Big Data & Society*, vol. 9, no. 1, p. 20539517221108370, Jan. 2022, doi: 10.1177/20539517221108369.

[24] N. Tuptuk, P. Hazell, J. Watson, and S. Hailes, 'A Systematic Review of the State of Cyber-Security in

Water Systems', *Water*, vol. 13, no. 1, Art. no. 1, Jan. 2021, doi: 10.3390/w13010081.

[25] R. S. H. Piggin and H. A. Boyes, 'Safety and security — A story of interdependence', in *10th IET System Safety and Cyber-Security Conference 2015*, Oct. 2015, pp. 1–6. doi: 10.1049/cp.2015.0292.

[26] K. Reegård, C. Blackett, and V. Katta, *The Concept of Cybersecurity Culture*. 2019. doi: 10.3850/978-981-11-2724-3_0761-cd.

[27] E. H. Schein, 'Organizational Culture and Leadership', p. 458, 1985.

[28] R. von Solms and J. van Niekerk, 'From information security to cyber security', *Computers & Security*, vol. 38, pp. 97–102, Oct. 2013, doi: 10.1016/j.cose.2013.04.004.

[29] F. W. Guldenmund, 'The nature of safety culture: a review of theory and research', *Safety Science*, vol. 34, no. 1, pp. 215–257, Feb. 2000, doi: 10.1016/S0925-7535(00)00014-X.

[30] A. da Veiga, L. V. Astakhova, A. Botha, and M. Herselman, 'Defining organisational information security culture—Perspectives from academia and industry', *Computers & Security*, vol. 92, p. 101713, May 2020, doi: 10.1016/j.cose.2020.101713.

[31] K. Dewey, G. Foster, C. Hobbs, and D. D. Salisbury, 'Nuclear Security Culture in Practice', p. 46, 2021.

[32] T. M. Bisbey, M. P. Kilcullen, E. J. Thomas, M. J. Ottosen, K. Tsao, and E. Salas, 'Safety Culture: An Integration of Existing Models and a Framework for Understanding Its Development', *Hum Factors*, vol. 63, no. 1, pp. 88–110, Feb. 2021, doi: 10.1177/0018720819868878.

[33] A. Zanutto, B. Shreeve, K. Follis, J. Busby, and A. Rashid, 'The Shadow Warriors: In the no man's land between industrial control systems and enterprise IT systems', p. 6.

[34] T. O. Naevestad, J. H. Honerud, and S. F. Meyer, 'How can we explain improvements in organizational information security culture in an organization providing critical infrastructure?', in *Safety and Reliability - Safe Societies in a Changing World*, S. Haugen, A. Barros, C. VanGulijk, T. Kongsvik, and J. E. Vinnem, Eds., Leiden: Crc Press-Balkema, 2018, pp. 3031–3039. Accessed: Feb. 28, 2022. [Online]. Available: https://www.webofscience.com/wos/woscc/summary/0c5f245f-6b80-49b6-8e6f-0a8b788258c4-267f1b99/relevance/1

[35] T. O. Naevestad, S. F. Meyer, and J. H. Honerud, 'Organizational information security culture in critical infrastructure: Developing and testing a scale and its relationships to other measures of information security', in *Safety and Reliability - Safe Societies in a Changing World*, S. Haugen, A. Barros, C. VanGulijk, T. Kongsvik, and J. E. Vinnem, Eds., Leiden: Crc Press-Balkema, 2018, pp. 3021–3029. doi: 10.1201/9781351174664-379.

[36] O. Michalec, S. Milyaeva, and A. Rashid, 'Reconfiguring governance: How cyber security regulations are reconfiguring water governance', *Regulation & Governance*, vol. n/a, no. n/a, 2021, doi: 10.1111/rego.12423.

[37] T. Wallis, G. Paul, and J. Irvine, *Organisational Contexts of Energy Cybersecurity*. 2021.

[38] J. B. Quinn, 'Strategic Outsourcing: Leveraging Knowledge Capabilities', *MIT SMR*, Jul. 1999, Accessed: Feb. 18, 2023. [Online]. Available: https://sloanreview.mit.edu/article/strategic-outsourcing-leveraging-knowledge-capabilities/

[39] S. Nevo, M. R. Wade, and W. D. Cook, 'An examination of the trade-off between internal and external IT capabilities', *The Journal of Strategic Information Systems*, vol. 16, no. 1, pp. 5–23, Mar. 2007, doi: 10.1016/j.jsis.2006.10.002.

[40] A. Bradshaw, V. Pulakanam, and P. Cragg, 'Knowledge Sharing in IT Consultant and SME Interactions', *AJIS*, vol. 19, Oct. 2015, doi: 10.3127/ajis.v19i0.1026.

[41] M. Pozzebon and A. Pinsonneault, 'The Dynamics of Client-Consultant Relationships: Exploring the Interplay of Power and Knowledge', *Journal of Information Technology*, vol. 27, no. 1, pp. 35–56, Mar. 2012, doi: 10.1057/jit.2011.32.

[42] A. Bradshaw, P. Cragg, and V. Pulakanam, 'Do IS consultants enhance IS competences in SMEs?', *The Electronic Journal of Information Systems Evaluation*, vol. 16, pp. 13–24, Jan. 2012.

[43] Y. M. Ha and H. J. Ahn, 'Factors affecting the performance of Enterprise Resource Planning (ERP) systems in the post-implementation stage', *Behaviour & Information Technology*, vol. 33, no. 10, pp. 1065–1081, Oct. 2014, doi: 10.1080/0144929X.2013.799229.

[44] D.-G. Ko, L. J. Kirsch, and W. R. King, 'Antecedents of Knowledge Transfer from Consultants to Clients in Enterprise System Implementations', *MIS Quarterly*, vol. 29, no. 1, pp. 59–85, 2005, doi: 10.2307/25148668.

[45] C. Cerruti, E. Tavoletti, and C. Grieco, 'Management consulting: a review of fifty years of scholarly research', *Management Research Review*, vol. 42, no. 8, pp. 902–925, Jan. 2019, doi: 10.1108/MRR-03-2018-0100.

[46] B. P. Bloomfield and A. Danieli, 'The Role of Management Consultants in the Development of Information Technology: The Indissoluble Nature of Socio-Political and Technical Skills*', *Journal of Management Studies*, vol. 32, no. 1, pp. 23–46, 1995, doi: 10.1111/j.1467-6486.1995.tb00644.x.

[47] L. Argote, B. McEvily, and R. Reagans, 'Managing Knowledge in Organizations: An Integrative Frame-

work and Review of Emerging Themes', *Management Science*, vol. 49, no. 4, pp. 571–582, 2003.
[48] J. Haney, W. Lutters, and J. Jacobs, 'Cybersecurity Advocates: Force Multipliers in Security Behavior Change', *IEEE Security & Privacy*, vol. 19, no. 4, pp. 54–59, Jul. 2021, doi: 10.1109/MSEC.2021.3077405.
[49] M. Gale, I. Bongiovanni, and S. Slapnicar, 'Governing cybersecurity from the boardroom: Challenges, drivers, and ways ahead', *Computers & Security*, vol. 121, p. 102840, Oct. 2022, doi: 10.1016/j.cose.2022.102840.
[50] A. Poller, L. Kocksch, S. Türpe, F. A. Epp, and K. Kinder-Kurlanda, 'Can Security Become a Routine?: A Study of Organizational Change in an Agile Software Development Group', in *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, Portland Oregon USA: ACM, Feb. 2017, pp. 2489–2503. doi: 10.1145/2998181.2998191.
[51] A. Bryman, *Social Research Methods*. Oxford University Press, 2016.
[52] V. Braun and V. Clarke, 'Using thematic analysis in psychology', *Qualitative Research in Psychology*, vol. 3, no. 2, pp. 77–101, Jan. 2006, doi: 10.1191/1478088706qp063oa.
[53] G. W. Ryan and H. R. Bernard, 'Techniques to Identify Themes', *Field Methods*, vol. 15, no. 1, pp. 85–109, Feb. 2003, doi: 10.1177/1525822X02239569.
[54] M. Williams and T. Moser, 'The art of coding and thematic exploration in qualitative research', *International Management Review*, vol. 15, no. 1, pp. 45–55, 2019.
[55] National Protective Security Authority, 'Think Before You Link (TBYL) | NPSA'. https://www.npsa.gov.uk/security-campaigns/think-you-link-tbyl-0 (accessed May 18, 2023).
[56] E. C. Page, 'Bureaucrats and expertise: Elucidating a problematic relationship in three tableaux and six jurisdictions', *Sociologie du Travail*, vol. 52, no. 2, pp. 255–273, Apr. 2010, doi: 10.1016/j.soctra.2010.03.021.
[57] European Parliament, 'The NIS2 Directive: A high common level of cybersecurity in the EU | Think Tank | European Parliament'. https://www.europarl.europa.eu/thinktank/en/document/EPRS_BRI(2021)689333 (accessed Aug. 15, 2022).
[58] W. Alkalabi, L. Simpson, and H. Morarji, 'Barriers and Incentives to Cybersecurity Threat Information Sharing in Developing Countries: A Case Study of Saudi Arabia', in *2021 Australasian Computer Science Week Multiconference*, Dunedin New Zealand: ACM, Feb. 2021, pp. 1–8. doi: 10.1145/3437378.3437391.
[59] D. Norris, A. Joshi, and T. Finin, 'Cybersecurity Challenges to American State and Local Governments', presented at the 15th European Conference on eGovernment, 2015.
[60] D. Ashenden, 'Information Security management: A human challenge?', *Information Security Technical Report*, vol. 13, no. 4, pp. 195–201, Nov. 2008, doi: 10.1016/j.istr.2008.10.006.
[61] J. M. Blythe, L. Coventry, and L. Little, 'Unpacking security policy compliance: The motivators and barriers of employees' security behaviors', presented at the Symposium on Usable Privacy and Security (SOUPS), 2015.
[62] D. Ashenden and A. Sasse, 'CISOs and organisational culture: Their own worst enemy?', *Computers & Security*, vol. 39, pp. 396–405, Nov. 2013, doi: 10.1016/j.cose.2013.09.004.
[63] D. Burrell, 'An Exploration of the Critical Need for Formal Training in Leadership for Cybersecurity and Technology Management Professionals', 2019, pp. 1420–1432. doi: 10.4018/978-1-5225-8356-1.ch069.
[64] I. Kirlappos, S. Parkin, and A. Sasse, 'Learning from "Shadow Security:" Why Understanding Non-Compliant Behaviors Provides the Basis for Effective Security', Feb. 2014. doi: 10.14722/usec.2014.23007.
[65] S. Schinagl and R. Paans, 'Communication Barriers in the Decision-making Process: System Language and System Thinking', presented at the Hawaii International Conference on System Sciences, 2017. doi: 10.24251/HICSS.2017.738.
[66] M. Bada, A. M. Sasse, and J. R. C. Nurse, 'Cyber Security Awareness Campaigns: Why do they fail to change behaviour?', p. 11.
[67] K. Li, K. M. Ramokapane, and A. Rashid, '"Yeah, it does have a...Windows `98 Vibe'': Usability Study of Security Features in Programmable Logic Controllers'. arXiv, Aug. 04, 2022. Accessed: Sep. 29, 2022. [Online]. Available: http://arxiv.org/abs/2208.02500
[68] J. C. Le Coze, 'How safety culture can make us think', *Safety Science*, vol. 118, pp. 221–229, Oct. 2019, doi: 10.1016/j.ssci.2019.05.026.
[69] R. L. D. Costa, N. António, M. Sampaio, and I. Miguel, 'The boundaries in the area of knowledge transfer in management consulting', *Gest. Prod.*, vol. 28, no. 1, p. e4956, 2021, doi: 10.1590/1806-9649.2020v28e4956.
[70] 'Human factors/ergonomics – Intelligent customer capability'. https://www.hse.gov.uk/humanfactors/topics/customers.htm (accessed Feb. 14, 2023).
[71] IAEA, 'Nuclear Security Culture', 2008. https://www.iaea.org/publications/7977/nuclear-security-culture (accessed Nov. 27, 2021).
[72] J. D'Arcy and G. Greene, 'Security culture and the employment relationship as drivers of employees' security compliance', *Information Management & Computer Security*, vol. 22, no. 5, pp. 474–489, Jan. 2014, doi: 10.1108/IMCS-08-2013-0057.

[73] M. Chan, I. Woon, and A. Kankanhalli, 'Perceptions of Information Security in the Workplace: Linking Information Security Climate to Compliant Behavior', *Journal of Information Privacy and Security*, vol. 1, no. 3, pp. 18–41, Jul. 2005, doi: 10.1080/15536548.2005.10855772.

[74] W. M. Cohen and D. A. Levinthal, 'Absorptive Capacity: A New Perspective on Learning and Innovation', *Administrative Science Quarterly*, vol. 35, no. 1, pp. 128–152, 1990, doi: 10.2307/2393553.

[75] S. A. Zahra and G. George, 'Absorptive Capacity: A Review, Reconceptualization, and Extension', *The Academy of Management Review*, vol. 27, no. 2, pp. 185–203, 2002, doi: 10.2307/4134351.

A Purdue Reference Architecture Model

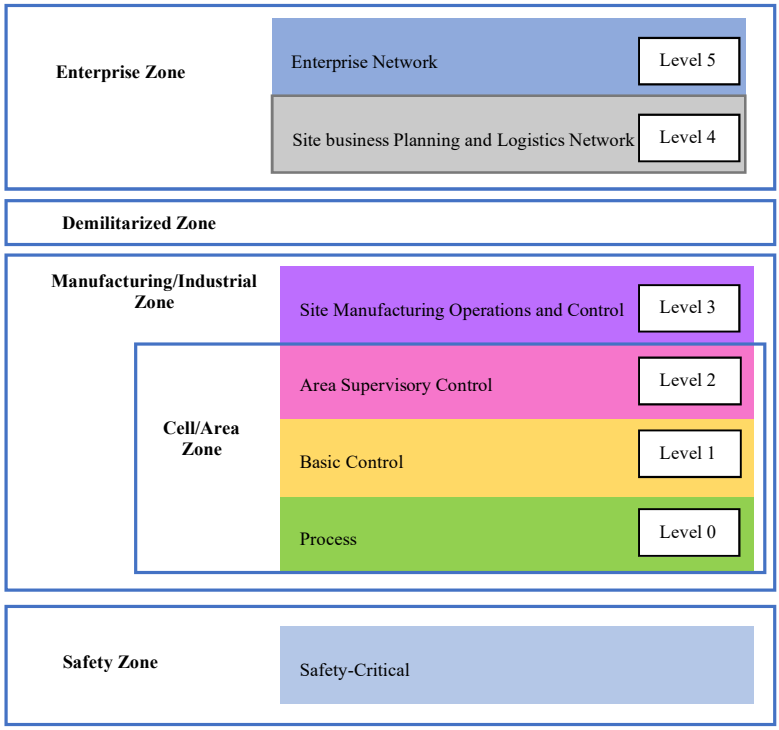


Fig. 1 Purdue Reference Architecture Model

B Participants' Role and Sector

#	Role	Sectors
P01	OT Manager	Water
P02	OT Manager	Water
P03	OT Manager	Energy
P04	Consultant	Transport
P05	Security Researcher	Manufacturing
P06	Security Manager	Water
P07	CISO	Energy
P08	CISO	Water
P09	Consultant	Transport
P10	Consultant	Various, p. Transport
P11	Security Manager	Energy
P12	Consultant	Various, p. Oil & Gas
P13	CISO	Transport
P14	Security Manager	Transport
P15	Security solutions vendor	Various
P16	Consultant	Various, p. Energy
P17	Security solutions vendor	Various
P18	Consultant	Various
P19	Regulator	Energy
P20	Regulator	Transport
P21	Consultant	Various, p. Manufacturing
P22	Consultant	Various, p. Oil and Gas
P23	Technology Manager	Water
P24	Government & Organisations Coordinator	Various, p. Maritime
P25	OT Manager	Energy
P26	Security solutions vendor	Transport
P27	Security Researcher	Various
P28	Academia & Industry Coordinator	Maritime
P29	Consultant	Various
P30	Consultant	Transport
P31	Consultant	Energy

P32	Consultant	Various
P33	Security Manager	Energy

Note: various p. signifies that a participant works across various sectors but primarily operates in one.

C Sample interview topic guide

We provide a sample interview guide that was used as a basis for our interviews with consultants and security product vendors.

- Participant's background
  - Previous & Current roles
- Details about company: services provided, structure of the company.
  - Can you describe the typical sales/consultancy process and/or a challenging case?
- In-depth questions about products, services
  - Relate to publicly available information (blogposts, websites, talks etc.) and previous conversations/emails.
- Security culture: How would you define it?
  - Who is responsible? Who else plays a part?
  - Challenges? What works?
  - If safety culture is mentioned – parallels, differences?
- Intra-organisational collaborations (if applicable)
  - Partnerships between vendors and consultancies,
  - Security solution providers' relationship with their supply chain etc.
  - Relationships with competent authorities
- Differences between sectors and/or companies
  - On how they utilise participants' services
  - Generally, with respect to their security maturity
- Debriefing & thanking for participation

D Codebook

We provide some key themes reported on this work, with a description and example quotes. Codes in bold are axial codes representing a wider theme, while those underlined are more specific open codes.

Governance and Management

Description: Instances of governance and management issues, or ways to improve this

Governance issues: *"As companies have organically grown over the years and I've typically seen this in all the oil and gas sector is, and in the electricity sector, departments become quite heavily siloed and fragmented, and they only talk up when they actually sit alongside each other and they're all actually doing something in a chain."*

Communication

Description: Instances of communication barriers, or ways to improve this situation.

OT and IT communication: *"And often a standing point where neither of them effectively communicates with each other."*

"We will be there sat in between the IT department and their engineering OT department trying to kind of translate between the two and get them to understand each other's point of view and this kind of stuff."

Communication with the board: *"I asked the NCSC to come and present to the board with me. So we had a kind of government lens and a … member of the security services stood in front of the board telling them there's a real risk. Really, really made people think, wake up and think."*

Expertise

Description: Instances of gaps in expertise, or ways to bridge the expertise gap.

OT Expertise: *"They often have an educational divide between cybersecurity teams that often don't understand operational technology and engineering teams who don't understand cybersecurity."*

"There is a talent shortage in OT cybersecurity at least for the energy sector. But my understanding is the energy sector is fairly mature, well, relatively mature and at the other sectors are further behind with the exception of maybe some oil and gas stuff."

General training: *"So one of the people from that team moved to my team and we've sponsored them on an MSc and so and that's the case. People who are more on the compliance side, where we're looking at NIS regulations, then they're looking more at sort of management of risk and qualifications or system, you know, the certificate in security management type qualifications. And then the security engineers that deal with operational technology, they tend*

to be more in the SANS space of the IEC 622443 type training. And so yeah it, as I say it really depends on the particular thing. And then you know intelligence analysts, we would use CREST training for, that would be the most relevant for them."

Security Culture

Description: Mentions of what a security culture entails, what falls under a security culture change process, or comparisons with other cultures

What is security culture: *"That cultural piece ... it's that understanding and inherent sort of guessing, ... you know you asked the right questions, you adopt the right behaviours, ... you design things securely, you make solutions that are secure."*

Comparisons with safety culture: *"I think security, if they've got a good safety culture it's not that much of a stretch for them to develop a good security culture on the OT side of things."*

Security culture and consultants

Description: Ways consultants contribute to an organisation, either by overcoming the three organisational obstacles or more generally how they shape factors that affect culture

Contribution of consultants: *"All about helping people who have OT plants to understand what their risks are. Usually start from absolute zero knowledge of OT security and help them get their head around where the gaps are, where the best place for them to spend their money is."*

Need for consultancy: *"There's a mix of reasons, so there are some ... companies that are really small. They have no security staff and now they're subject to NIS and then they can't, can't just build the team out of nowhere. Right. You have to delegate almost all responsibilities in other instances. It's just that the expertise isn't there."*

Security culture and security product vendors

Description: Ways security product vendors contribute to an organisation, either by overcoming the three organisational obstacles or more generally how they shape factors that affect culture

Contribution by vendors: "*We have to educate and bring people along with our way of thinking and understanding. ... Umm, but we have to be successful, get people asking the right questions.*

"We were primarily currently selling products of course, ... so I suppose from that from that sense, you know we're enabling this in that the alerts we produce and the data we can produce for companies, give them a window into their networks that they didn't have before. So knowledge, one of the foundations of culture is knowledge."

State of OT cybersecurity

Description: General comments on the state of OT cybersecurity by practitioners

State of OT cybersecurity: "*The cybersecurity industry is starting; they are babies at this game."*

"It's embryonic within the rail sector."

Important triggers for increased cybersecurity awareness: *"It's also the other last thing I would say with that as well is a rise in ransomware, is a massive concern as well, and we know that you know, for every ransomware attack that makes a headlines, there's many, many more that don't, because companies just end up paying it because they don't want the bad publicity."*

Lacking the Tools and Support to Fix Friction: Results from an Interview Study with Security Managers

Jonas Hielscher, Markus Schöps, Uta Menges, Marco Gutfleisch, Mirko Helbling, and M. Angela Sasse
Human-Centred Security, Ruhr University Bochum

Abstract

Security managers often perceive employees as the key vulnerability in organizations when it comes to security threats, and complain that employees do not follow secure behaviors defined by their security policies and mechanisms. Research has shown, however, that security often interferes with employees primary job function, causing friction and reducing productivity – so when employees circumvent security measures, it is to protect their own productivity, and that of the organization. In this study, we explore to what extent security managers are aware of the friction their security measures cause, if they are aware of usable security methods and tools they could apply to reduce friction, and if they have tried to apply them. We conducted 14 semi-structured interviews with experienced security managers (CISOs and security consultants, with an average 20 years experience) to investigate how security friction is dealt with in organizations. The results of the interviews show security managers are aware that security friction is a significant problem that often reduces productivity and increases the organization's vulnerability. They are also able to identify underlying causes, but are unable to tackle them because the organizations prioritize compliance with relevant external standards, which leaves no place for friction considerations. Given these blockers to reducing security friction in organizations, we identify a number of possible ways forward, such as: including embedding usable security in regulations and norms, developing positive key performance indicators (KPIs) for usable security measures, training security managers, and incorporating usability aspects into the daily processes to ensure security frictionless work routines for everyone.

1 Introduction

Security experts often describe humans as the key vulnerability in organizations [59, 69]: employees who are not aware of security threats, and do not follow prescribed secure behaviors. Usable security research established in the late 90s' [82] showed that in the contexts of employees' work goals and environment, their behavior is completely rational: they are hired, assessed and rewarded for performance on their primary job, so security policies and unusable security mechanisms that get in the way cause friction, and too much friction leads to security being circumvented [13]. Furthermore, friction does not only cost productive time and reduces innovation [47], it also makes organizations more vulnerable.

Even though some research has been done on investigating security tools for employees [19, 26, 29, 63, 91], only a few studies investigated usable security, and consequently also security friction, within real-world organizations [3, 18]. Security in companies cannot be achieved if the context in which employees find themselves is ignored. We therefore want to investigate how usable security and more specifically security friction are handled in organizations. This includes, perceptions of decisions makers, as well as consequences and causes of security friction. Therefore we focus on the following research questions:

Q1: How does organizational security management perceive security friction and deal with it?

Q2: What are the perceived causes of friction in organizations and its impact on the organization and its employees?

We reached out to highly experienced security managers and conducted $n = 14$ semi-structured interviews, 7 with CISOs and 7 with senior security consultants. Each interview focused on capturing their experiences and perceptions of security friction within organizations they worked for. With an average of 20 years of industry experience in large-scale organizations headquartered in a German-speaking region, we have addressed a wide variety of perspectives, measures and decisions made in organizations within the interviews and, consequently, also in our analysis.

Almost all our participants were aware about security friction and its relevance in the context of creating or implementing new security routines and policies. However, they described many cases, where usable security and hence resulting friction is not considered at all. As reasons for this,

they have cited both a lack of resources or strict external, as well as internal regulatory requirements. Additionally, caused friction is almost not measured and the reduction of friction, which might lead to boost in security and an increase in productivity, is not either. The active inclusion of friction in the decision making process and its consequences for productivity and security could be a first step towards more usable security.

To the best of our knowledge we are the first investigating security friction through the lens of security management within real-world contexts – which is also the case because especially CISOs have a busy, high-pressure job, so researchers asking for in-depth interviews face a challenge. Previous work focused either usable security within software engineering [39] or on end-users [55, 56]. Our contributions are the following: (I) we describe possible causes of frictions and the real world impact. (II) We highlight how security friction is perceived by our participants and how this shapes their security decisions within organizations. (III) We discuss open challenges in academia and give recommendations for industry and regulation authorities, how to establish more usable security routines and practices within organizations, e. g., that usable security should become embedded in regulations, norms and the security process [39, 41], that positive key performance indicators (KPIs) should be developed that highlight the (monetary and intellectual) savings that come with usable security measures, that security managers need to be actively trained in usable security, and that usability aspects should become part of procurement processes.

2 Related Work & Background

Here we summarize previous research about security friction in organizations (Section 2.1), as well as with and about security managers (Section 2.2).

2.1 Security Friction

Usable security research has the main goal of reducing the effort to use a secure tool or procedure [34] – explicitly and implicitly – and to increase the adoption rate of such [18]. Time and subjective satisfaction of the users is what needs to be achieved [34, 85]. While usable security studies often focus on understanding the (un)usability or improvement of tools, in our work we took a wider look: we consider *security friction* as a problem created through a multitude of badly written security policies and measures that cost time, effort, and nerves, and are not aligned with employees routines, ultimately leading to reduced productivity [76], shadow security [55] (the implementation of alternative security mechanisms by employees, if they perceive the prescribed as too complicated), or a reduced security level.

Herley [45] points out that security professionals often assume that employees only need to be convinced and persuaded to invest more time and effort in security, implying that employees would misjudge the cost-benefit trade-off, which has been refuted in most cases, for example by the concept of the compliance budget: the lack of adaptation of security and business processes leads to security friction, which, according to Beautement et al. [13], is the key to individual compliance problems. The compliance budget (consciously or unconsciously weighing the costs against the benefits) is further reduced when friction-triggering tasks accumulate or repeat, which can lead to employees no longer adhering to security guidelines. Blythe et al. [15] made it clear that managers in organizations are obliged to ensure that security rules are designed in such a way that they do not hinder the actual work.

In the context of security friction, the concept of *security fatigue* is notable. This phenomenon is described by Furnell [33] as a situation in which users, and thus employees, become tired of dealing with security and associated warnings. Various factors can trigger security fatigue, including the complexity of security tasks, constant confrontation with security measures and more. With regard to security friction, it was observed that employees feel security fatigue due to a state of friction between the fulfillment of security measures and primary job requirements and the resulting conflict [20]. Cram et al. [20] found, for example, that security fatigue can, among other things, lead to employees behaving in a risky manner when using computers in both work and private contexts. Furthermore, security fatigue should be considered as one of the costs users (employees) face when they are inundated with security rules [86].

In a two-fold study with 290 employees, Mayer et al. [62] found that productivity goal setting (KPIs) decreases security compliance – the goal to be productive being in direct conflict with following security policies. Albrechtsen [4], found that users fear a conflict of interest between *functionality and information security*. Molin et al. [67] recognize that CISOs should put personal productivity into consideration when putting security measures into place.

2.2 Security Managers

Some studies in the past looked at security managers, especially on CISOs, and investigated their role descriptions, tasks and backgrounds. While, to the best of our knowledge, no previous study looked at CISOs' perception of security friction, their role and the problems they are facing were part of some evaluations [11]. CISOs can mainly be found in larger organizations, while it is not strictly defined to whom they have to report [2, 27, 83]. Most CISOs have a background in computer science or engineering [28]. However, the required skill set of CISOs also includes *IT security skills* to defend, monitor, and protect [12, 49, 54, 93], *strategic security management and government* [8, 12, 38, 49, 54, 64], *leadership and communication skills* [8, 49, 93], and *security teaching skills* [8, 12, 54, 93].

Independent of their tasks, CISOs are under immense pressure and experience unhealthy levels of stress [70]. The experiences and opinions of CISOs and other security managers have been studied previously with regards to their security experiences in small and medium-sized enterprises [31, 50], their security budgeting decisions in agreement with the management [68], and their perceived role and collaboration in their organization [5, 9, 21–23, 30, 48, 60, 74, 77].

We are not aware of interview or questionnaire studies with a focus on security consultants – with the possibly closest studies being carried out with security advocates [42–44].

3 Method

We performed in-depth, semi-structured interviews with $n = 14$ highly experienced security professionals in highest security management positions to learn about their perception and handling of security friction in their organizations/ the organizations they advice. By combining the perspective of CISOs that drive security decisions from within the organization and security consultants, whose target groups are CISOs and high-level management, we are able to get internal and external perspectives on the topic. Our sample is small – while this specific population is in general rather small –, but they offer unique insights into incentives that drive decisions that create or prevent security friction. The interviews were organized as virtual conversations. They were carried out from April to June 2022. Our method is summarized in Figure 1.

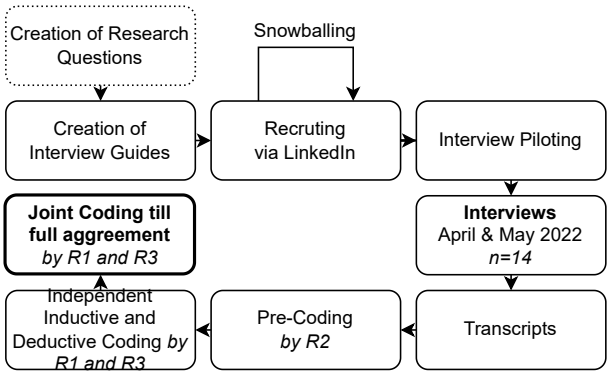


Figure 1: Our methodology.

3.1 Instrument Development

Within the following section we describe the structure of our interview guide and how it was developed. From previous research we can not assume that our participants have a uniform understanding of *usable security* and *security friction*. We therefore centered the interview guide around the organizations' employees in the context of security measures and decisions. Furthermore, within the first part of the interview guide we focused on understanding open challenges and their economic perspective. However, we only deal with these topics in our work if they were directly related to security friction.

Two interview guides were developed for both cohorts with slight differences: the questions for the consultants were asked around the organizations they advise, while the CISOs answered for their own organizations. The interview guides were developed by 4 researchers in multiple iterations over the course of 3 months. Due to the limited literature that focuses on security managers, only some guiding questions could be developed based on it, namely the questions around the relationship between employees and security managers [10, 22, 48]. One week before the first regular interview, we piloted our instrument with a security consultant, who gave feedback about the (I) administration, (II) interview atmosphere, (III) comprehensibility of the questions, and (IV) the content. Slight adjustments were made to the interview guide and the pilot interview was not included in our analysis.

Ultimately, all questions were organized around 8 guiding questions (see Appendix A for the full interview guide): firstly, we asked about the (personal) experience with security in the industry and their education. This was followed by questions about the biggest security challenges. The remaining six questions addressed employees' work routines, friction measurements, primary task conflicts in the organization and negative reactions from employees, as shown in Figure 2.

3.2 Recruitment

Since we aimed for highly experienced participants in management positions (and in larger organizations), we applied the following selection criteria: (I) participants had to be currently working as CISOs or (senior) security consultants, (II) they had to have at least 8 years of experience in the field of security, and (III) they had to either be qualified through an academic degree or relevant professional training (e. g., CISSP, CRISC, CISM) [72]. The recruitment happened in two steps: firstly, participants with according job titles, focused on the largest private and public organizations based in the country of our study, were searched on LinkedIn (33 in total, from which 10 did respond). In a second phase, the participants were asked whether they could provide other interesting interview partners (snowballing), which resulted in the recruiting of another 8 contacts. In the end, 14 interviews took place. We recruited in a German speaking country in Europe. The participants were native German speakers and were interviewed in German. We did not compensate the participants, because we did not feel that any monetary offer we could make would reach the hourly salary of the participants. Instead we offered to share our results.

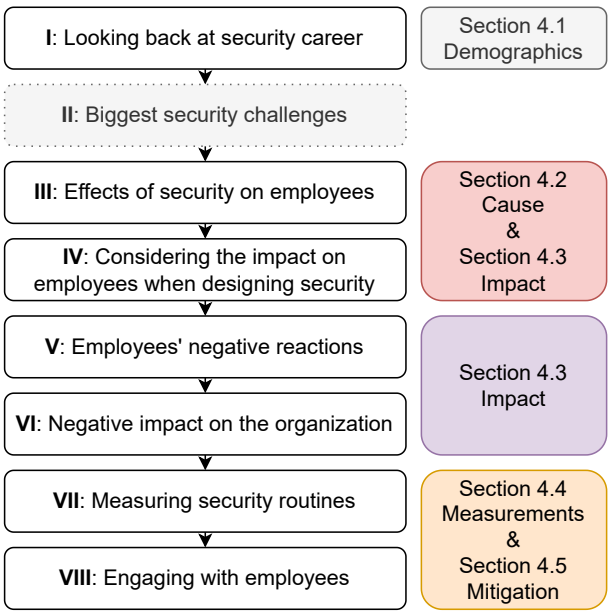


Figure 2: The 8 guiding topics (questions) of our interview guide(s), mapped to results in Sections 4 that mainly (but not exclusively) contain the answers.

3.3 Analysis

We applied Kuckartz' [58] process scheme of content-structuring analysis, combining deductive and inductive coding strategies and a category-based evaluation along main codes. The coding was done with MaxQDA and happened in multiple steps, carried out by three researchers (R1-R3) – all experienced coders – with two more (R4-R5) participating in the final analysis of the data: (I) in a first pre-coding step, R2 coded all 14 interviews to identify potential key topics. (II) Following this, R1 and R3 independently created deductive codebooks based on the interview guide and the research questions. (III) R1 and R3 then coded 5 different interviews (R1 coded 3, R3 coded 2) deductively and inductively. (IV) The codebooks were merged, reduced and superordinate key codes were identified. (V) One interview was deductively coded by both researchers, based on the merged codebook. (VI) We refined the codebook again and coded all remaining 8 interviews in a joint session, **until full agreement was reached**. (VII) In a final step R1,R3,R4 and R5 discussed the results of the coding process and how to present the results, which happened in multiple in-person and virtual meetings.

During all steps, multiple memos were created, guiding the discussion and analysis. Although we did not apply a saturation criteria to our sampling strategy, we experienced saturation during our analysis, as we found a high degree of overlap and repetition within the categories. Since the interviews were done in German, we translated those parts of the transcripts that we cited in the paper into English. The full codebook can be found in the Appendix C.

3.4 Ethics & Data Privacy

Our institution does not have an institutional review board (IRB) nor an ethics review board (ERB) for security research. We followed best practices in human subject research [89] and considered the deanonymization of the participants as the primary threat. We followed European data privacy guidelines (GDPR) and informed the participants about the study procedure and their rights prior to the interviews. All participants gave their agreement. As soon as the transcription of the audio files was completed, we deleted the audio files. We removed personal identifiers like names of individuals, organizations or other terms that might reveal the participants' identity. Furthermore, we kept the participants' country of residence, as well as the company they were working for a secret. We report some demographic data only in a pseudonymized or aggregated form. The community of CISOs is rather small and otherwise the demographic data we report here might reveal their identities.

3.5 Limitations

As with every study with human subjects there are several limitations in this study: all 14 participants were male. This is not only based on the fact that non-male security managers are underrepresented in the country of our study, but also due to the fact that we recruited through snowballing and the participants only suggested other male interview partners – a phenomenon well known as male-only-circles. The participants all present years of experience, with no one being new in this field. While one would expect this of a management position, this might have biased the results towards ignoring recent trends, only perceived by newcomers. Our study was performed in a European country, not all phenomenons might be found in other countries or cultures, e. g., due to different legislation. Given the challenge of getting enough time for an interview, we focused on a region where one of the authors had access to, and the trust of the participants. While they gave us a deep view into the security managers' perceptions of security friction, the results can not be generalized to a greater population.

4 Results

We first provide more details on our study participants (Section 4.1), before presenting our results about the causes of friction (Section 4.2), its impact on employees and the organizations (Section 4.3), how friction is measured (Section 4.4) and mitigated (Section 4.5), and, finally, how our participants perceive usable security (Section 4.6). Statements of CISOs are marked as *Ci1-Ci7*, those of consultants as *Co1-Co7*.

4.1 Demographics

We did ask biographic questions – about the educational background and experience – in the interviews. All 14 participants identified as male. Table 1 shows the most important demographic properties. To keep the participants anonymous, we do not report the exact education or years of experience, but the accumulated numbers in Table 2. The interviews lasted between 23 and 46 minutes, with an average of 34 minutes.

4.2 Causes of Friction

Within the following section, we describe participants' directly or indirectly mentioned causes of friction.

(Regulatory) Security Requirements Eight participants (Ci1, Ci2, Ci4, Ci5, Ci7, Co2, Co3, Co5) stated that following regulatory security requirements cause or can cause security friction: *"If the law makes it mandatory, then you have to do it, even if the employee is not entirely happy with it."* — [Ci2]. Furthermore, Ci2 expressed that there is no *debate* about whether to fulfill regulatory requirements or not. Ci4 explained that audits are so important that there is no room to consider friction: *"We also have our audits and therefore some things we just have to implement."* — [Ci4]. Especially if the focus is only on achieving a security certification, the implementation can suffer and cause friction: *"That leaves ISO27001 again. Because it is the most established. [...] The goal is: 'I want this certification.' And, to put it bluntly, you're almost walking over dead bodies. So now you [the employees] have to do it like this."* — [Co2]. Regulatory requirements, however, do not only come from the outside. Co5, who is working in the defense industry, explained that internal security policies are so strict that he only can support the employees to a certain extent: *"There is an attempt to provide employees with as many aids and assistance as possible. But it certainly cannot be taken into account to the same extent as perhaps in other places."* — [Co5].

New Security Around half of our participants explained that the introduction of new (stronger) security policies causes employees' disapproval: *"And in the worst case employees take this negatively, because something that used to work well then doesn't work anymore."* — [Ci3]; *"And then the bad guys are the security people, because they now demand something that wasn't necessary before, and that is the reason what leads to these backlashes."* — [Ci3]. The security managers view security as an additional expense for the employees in most cases. For example, Ci4 explained that participants have to do a lot more steps, because Multi-Factor Authentication (MFA) was introduced, and it is considered *normal* that participants react negatively, as they have to do more than before. Co1 reported that restrictions are often put in place before workable solutions/ alternatives are implemented: *"Implementing negative measures before you have the positive benefits. So banning messaging tools before you have a tool that is acceptable security-wise. So it doesn't matter now if we use Instagram, WhatsApp, or something. If I ban it for security reasons, then that produces negative reactions."* — [Co1] He added that the friction grew following more restrictions after a public security agency published warnings: *"Until last year, we still had the possibility to receive older office file formats via various channels, because they are still in circulation. [...] this led to an emergency change in the fall, so very quickly stricter restrictions were introduced on various channels."* — [Co1]

Lack of Resources Some participants stated that they did not have sufficient resources (money or time) to consider friction. Ci3 reported that the reduction of friction is possible, but that it comes with a cost: *"It is also relatively often possible to make this comfortable for the employees. The problem is that it costs money. Cheap measures are often taken at the expense of the employees."* — [Ci3]. Furthermore, Ci3 mentioned that if not enough resources are available, *"a lot has to be solved via guidelines or instructions"* — [Ci3]. Co1 also stated that the advantages of low-friction solutions can not (easily) be monetized, in difference to security awareness (trainings): *"If we do awareness, then every employee now has to do an e-learning in, let's say, 30 minutes. 30 minutes at an hourly rate anyway. That time costs. [...] What is charged less are, for example, the improvement possibilities. So let's take implementation of an identity and access management system. Instead of having to manage 20 passwords, for X systems, or so, I only have one password. One central authentication. That effectively gives savings. But you can't monetize that. Or very difficult to monetize."* — [Co1].

Old and Poorly Designed Security Two participants (Ci3,Ci7) described that old products, routines or services slowed down the implementation of modern (more usable) security structures or mechanisms: *"it will probably go on for some time until virtually all the legacy that we have built up over the last thirty years is somehow no longer there [...]"* — [Ci3]. But also bad designed awareness campaigns, or poorly planned security initiatives might cause friction and reactance: *"Poorly designed security measures or poorly designed awareness campaigns always lead to resistance at the beginning."* — [Co1].

Short Summary: Our participants perceive regulations and norms as a major cause of friction – as they do not believe that these leave room for taking employee demands into account. They also report that the introduction of new security policies and measures create friction.

Table 1: Demographic information of our participants. *Experience* is the experience in the field of information security in years. *Origin* describes the first touch points the participants had with information security. *Usab.* (Usability Importance) shows the participants answers to the question: "How important do you rate the issue on scale from 1 (not important) to 10 (very important) that security measures can be integrated into the work routines of employees?". Some participants decided to answer outside the scale with 11. *Size* is the number of employees of the organizations the CISOs are working for. *Dur.* is the duration of the interviews in minutes.

P	Sector	Education	Experience	Origin	Usab.	Size	Dur.
			CISOs				
Ci1	Public Sector	Certificates	20-25	Security Revision, Consulting	9	>30,000	25
Ci2	Finance	Master/ Diploma	20-25	Consulting	10	?	27
Ci3	Transportation	Master/ Diploma	10-14	Technical IT Security (Firewalls, etc.)	10	>30,000	31
Ci4	Finance	Vocational Training	>25	Mainframe IT	11	1,200	56
Ci5	Insurance	Master/ Diploma	>25	Organizational Security	7-8	4,500	43
Ci6	Construction	Master/ Diploma	22-25	Cryptography	10	3,500	22
Ci7	Banking	Master/ Diploma	20-25	Technical IT Security	10	900	32
			Consultants				
Co1	Consulting	Master/ Diploma	>25	Security Revision, Consulting	11		34
Co2	Consulting	Master/ Diploma	20-25	Penetration Testing	11		37
Co3	Consulting	Master/ Diploma	10-14	Technical IT Security	7-8		23
Co4	Consulting	Master/ Diploma	15-19	IT Administrator	10		46
Co5	Defense	Certificates	5-9	Project Consulting	9-10		27
Co6	Consulting	Certificates	N/A	Business Continuity Consulting	10		30
Co7	Consulting	Certificates	10-14	Politics Advisor	8		40

4.3 Impact of Friction

The security managers described the negative effect of security friction on employees, and the organization as such, which we elaborate in this section.

Circumventing Security Eight security managers (Ci1,Ci2,Ci5-Ci7,Co2-Co4) described that, in order to avoid friction between their primary task and the organization's security measures, the employees would seek for opportunities to circumvent the organizational security measures. Often it is about saving time to complete work tasks faster (and more comfortably), e. g., in regard to the handling of data: *"[...] that you send things home because it's so wonderfully convenient. Or that there is an instruction that data must be sent by web transfer, but you still send it unencrypted by e-mail because it's just faster."* — [Ci1]. The danger that was explicitly pointed out was that security measures can increase security to a certain extent, but that this could also have the opposite effect: if the employees are dissatisfied or feel disturbed by the security measures, it can happen that the security is circumvented, which creates new risks (*"The measures are circumvented. The security instructions are not followed. This creates new risks."* — [Ci6]). Another interviewee made it clear that reactions that express the experience of friction must be dealt with appropriately; and that it is precisely these *negative* reactions that are highly relevant. Especially regarding incomprehensible and impractical measures, there would be reactance and defensive reactions and ways of circumventing them. Furthermore, it was discussed that practical resistance and circumvention possibilities get around and are thus quickly spread.

Negative Reactions Triggered by Friction The interviews reveal a variety of possible reactions of employees to security measures that generate friction. As concrete examples, two interviewees (Ci2, Ci3) mentioned the development of *shitstorms* – an accumulation of incomprehension, frustration, displeasure, etc. – in response to new, changed and poorly implemented measures that can *escalate*, especially when several people in the team feel affected. In addition to possible resignations of employees, Co4 described the following form of employee reaction: *"[...] making a fist in your pocket sometimes and just up to refusing to work or just doing the exact opposite"* — [Co4]. Declining motivation or anger, triggered by a feeling of being overwhelmed, are also described as reactions. According to Ci3, negative feedback would be received especially if the security department had gone too far.

Restricting and Work-Impeding Security The respondents mainly described that friction for employees is that their work becomes *more difficult* due to security measures. This means that processes take longer and are more cumbersome, and that goals are achieved more slowly. Security measures are perceived as creating friction when restrictions are the result: *"You can no longer do everything as an employee. You are no longer available and so on. In other words, increased security typically leads to restrictions in the first phase, which*

are perceived negatively" — [Co1]. Other examples of restrictions are the prohibited exchange via cloud platforms or the use of flash drives. Examples of tedious and work-inhibiting processes are requirements for long passwords, frequent entry of a second factor or a screen time-out after five minutes.

Some of the interviewees clearly stated that employees want to do their main work (*primary task*) [37], also because they are measured by their achievement of goals and productivity. Security (*secondary task* [25,82]) means the investment of time that is lacking elsewhere. Ci4 gave the example of a nurse who might have such conflicts: *"Explain to a nurse, and I certainly have a very high level of understanding that she has to lock the screen of the departmental PC [...] And has to unlock it again when she comes back to the PC. The argument that they hear that can cost lives. Because it costs time. Yes, but it can also cost lives or at least have a bad impact on lives if someone wrong has access to the data."* — [Ci4].

Endangering Economic Efficiency Also (partly potential) economic effects were described in connection with friction through security measures. An increasing fluctuation rate of the employees due to emerging frustration about that processes are too slow, hinder work or block the achievement of defined project goals, was mentioned as an example. The output that employees could provide and the organization's revenue that depends on it can be negatively affected by friction-triggering security measures, such as the multiple entry of passwords. Employees may be discouraged from performing the associated work tasks regularly or need more breaks. Co7 has been clear about this: *"[...] we are shooting ourselves in the foot if we make employees there dissatisfied and also negatively influence the profitability of an organization"* — [Co7]

Short Summary: The security managers described various effects triggered by friction, such as the circumvention of security measures, which can lead to new risks, or the deterioration of the quality of relationships between employees and the security department. Resignation, frustration and decreasing motivation were described as reactions of the employees, which in turn can negatively influence the economic efficiency of the organization.

4.4 Friction & Routine Measurements

Importance of Friction Measurements In general, all security managers (except Co4) did provide ideas about how to measure security friction or get insights into employees routines. Even though most managers describe friction measurements as being important, some (Ci2,Ci6,Ci7, Co1,Co3,Co5-Co7) mentioned that it is often not followed through in the organization (*"In my experience, far too little. You do something, you can check it off. Okay, we've now implemented another heuristic spam filter. Check it off. And then the job is done. You don't ask, has it gotten better now?"* — [Co1]), or that measurements have little impact on the implementation of measures (Co1,Co2,Co5): *"Of course there is the possibility to give feedback. But I think that usually it has little influence, because it's just the guidelines and has to be adapted that way."* — [Co5].

The most frequently mentioned measurement method was to simply talk with and listen to employees (Ci1-Ci2,Ci4-Ci7,Co1-Co3,Co5-Co7). Other examples were target group analyses (Ci1,Co1,Co3), surveys (Ci5,Co3) or technical measurements (Ci5-Ci7,Co1). These types of measurements were not all viewed positively: two managers mentioned the advantage of technical measurements compared to surveys (Ci6,Ci7), with one relying strongly on these measurements *"[...] many people are in the home office, okay? so you don't actually have to start a survey. We have very clear key figures from the systems."* — [Ci7] One manager described using technical methods to measure friction by monitoring employees' rule breaks *"I see on the one hand, yes, issues when people have trouble implementing something because it doesn't work or because it's difficult or something. And I can also detect rule violations and so on in a technical way."* — [Ci6] One manager directly stated the necessity of measuring before implementing security mechanisms (*"Before I instruct or regulate anything, I first have to understand what people are doing so that I can evaluate whether what I am proposing makes any sense at all and fits in with it."* — [Ci1]) following that the consequence of not doing so might lead to the non-acceptance of employees: *"[...] and if you don't do exactly that, then you won't have any understanding from your employees."* — [Ci1]. On different occasions managers (Ci4,Ci7,Co1,Co6) mentioned the importance of considering employees' wishes, even if it meant hearing negative reactions *"A negative reaction means that someone dared to react and these reactions are particularly important."* — [Ci7]

Obtaining Direct Feedback Some managers (Ci3,Ci7) mentioned a kind of "distance" to the employees, which resulted in only superficial measurements of friction: *"We also like it when comments come in unfiltered. That's what I said at the beginning, because it's very hierarchical, you sometimes don't feel the pulse of the employees."* — [Ci3]. Other managers (Ci4,Ci5,Co1,Co2,Co6) mentioned the importance of being close to employees to get an accurate view of whether the implementation of security measures worked and if they were accepted: *"You see, the only thing that helps there is proximity to the base. You have to somehow manage to get feedback from the employees as to whether the measures can be implemented, whether the measures are credible."* — [Co6], sometimes highlighting casual situations as the best way of getting feedback: *"[...] as a security officer, I have to get out among the people. And talk to them. At company meetings, company events. Departmental events. Lunch. Whatever."* — [Co1]. Another manager highlighted empathy as a necessary character trait of the person measuring: *"He*

has the right methods, but he may not have the right empathy. He doesn't have the understanding of the process. He doesn't have the understanding of the interplay, the interlocking on the human side, but also on the technical side. And that's the thing that it takes for the measures to be accepted." — [Ci4].

Short Summary: Most security managers are aware of the importance of friction measurements and employees' feedback, often citing casual talks as the best way of doing so. Still, this is not followed through and measurements are described as having no impact on security decisions.

4.5 Friction Mitigation Strategies

Different mitigation strategies – to reduce friction – were named by the security managers. However, we found that the majority of participants did not consider such reduction before (the interview) and did not name concrete examples where they applied strategies that would eliminate the causes of friction.

Awareness Will Solve Friction By far the most frequently presented mitigation strategy (named by 12/14 managers) was the idea to explain the importance of security and why restrictions are necessary to employees so that they will accept them and stop complaining. Some security managers insisted that a pro-active and open communication with the employees is key to raise their understanding: *"Security is [...] perceived as somewhere, maybe an obstacle or something. So we're aware of that, and we try to maintain a positive image, i.e. that people can approach us at any time. But security has priority, of course."* — [Ci7]. The majority of managers combined their suggestions with a form of excuse: they would not be in the position of bending rules and norms and hence can not do anything to reduce the friction: *"So the supreme law and regulation, what does it say? If the law requires it, then you have to do it, even if the employee is not entirely happy with it."* — [Ci1]. Others stated that, especially in their industries, the employees need to understand why security is so strict and causes problems: *"I think that's also primarily a mental attitude that has to take place that we're in a company that doesn't function like we do at home, because we're operating in very sensitive areas."* — [Ci2]. Some gave concrete examples where they would use explanations to solve the problem: *"But at best, if an employee is dissatisfied that a longer password than before suddenly has to be entered, then you simply have to communicate that."* — [Co7].

Develop Security Together Five security managers (Ci1,Ci2,Ci3,Co4,Co5) in some form or another expressed the idea to adapt security in collaboration with the employees. This ranges from implementing actively gathered feedback to the inclusion of the employees in the security requirement engineering process: *"And consequently, via data classification and categorization and protection needs analysis, it is then clear how much and where protective measures must be applied that then just together with the users, must also be balanced."* — [Co4].

Change Security Four security managers (Ci2,Ci3,Co1,Co3) were open to changing security/ lowering the level of security to reduce friction. Co3, for example, explained that security policies must be bent if the job requires it: *"[...] need to get changed, if that is too restrictive, or incompatible with the field of activity. Just as a sales person is often on the road, probably needs flash drives to work and exchange data. And to forbid him to do so would probably be bad."* — [Co3] Co1 explained that, over time, it is possible to consolidate security policies which would also reduce friction: *"[...] products have been standardized, harmonized, guidelines have also been slimmed down. And so on. So, if security is at a high level. Then it effectively becomes easier for the employees. And not more difficult."* — [Co1]

Others Only one manager said that he would help employees to practically train the security procedures to reduce friction, in that case with the unsolved usability problem of e-mail encryption [78, 84, 92]: *"A typical example is sending confidential information by e-mail [...] what exactly does he have to do? What is the button in the e-mail program where I activate the encryption? How can I see that it's all working? Things like that."* — [Ci1] Another idea was to offer secure alternative software to replace those that employees are used to, but are banned for security reasons. Co1 was convinced that all messengers are the same, and that it would be easy to introduce a secure messenger as an alternative for a more insecure but popular one (like WhatsApp): *"If I offer a messenger that allows end-to-end encryption and allows confidentiality in the relationship. Then there's no negative reaction because now I might have to switch from product A to product B. But I can communicate."* — [Co1]

Short Summary: Most security managers propose mitigation strategies that rather hide the friction but do not solve it – namely the idea to convince the employees that restrictions and friction are necessary in the name of security.

4.6 Usable Security

The managers hinted at an understanding of usable security. While the term *usable security* was not used by any manager, *usability* was mentioned 5 times by 4 managers (Ci1,Co3,Co5,Co7) and the term *user friendly* 8 times by 4 managers (Ci3,Ci6,Co4,Co7), e. g.,: *"Legitimate is certainly the desire for usability that you do something securely, but*

not so complicated that it takes away a significant amount of work time that it's understandable that the employee has a sense of security in what they're doing that they're doing it right." — [Ci1]

The managers mentioned software and mechanisms that could make security usable, namely password managers, Single Sign-On (SSO), biometric authentication and MFA codes on mobile devices, e. g.,: *"And basically the subject of single sign-on. If I want or have to log on to different platforms because I need different tools, different applications, different services, but can largely cover this with SSO that's an increased security feature. But at the same time an improvement in user-friendliness."* — [Co7]. However, while multiple mechanisms were named by the participants, they always talked about them in abstract forms, never mentioning that they had introduced such themselves in their organizations. The only exception was Co3 who gave an example about how he implemented a usability concept: *"I found that the screen timeout is set to something like two hours and the settings are not up to security standards at all. [...] The people who work on these systems sometimes wear gloves, there are three or four screens around this machine and that would definitely not be usable or compatible with today's standard rules if the screen saver came on every 15 minutes without anything being pressed."* — [Co3]

Invisible Security The idea to make security invisible to the employees – a concept that the usable security community can not agree upon to date [24] – was brought up by some managers: *"So in the best case, not at all. So if we, let's stay with the example of user authentication, if that goes by very gently, so that we don't virtually burden the the employee with security, but rather check that in the background."* — [Ci3] or *"Before disk encryption was introduced, you had to find a tool. Find a solution. Which makes this very transparent in the background, without employees noticing or feeling it. You switch it on and at some point, over the next few hours or days, it will be encrypted in the background. Such a security measure is accepted. Because it doesn't affect users, hinder them."* — [Co1]

Problematic Understanding of Usable Security Ci6 showed quite a controversial understanding of usable security. Employees told him that a security mechanism does not work on mobile devices and instead of improving the UI/UX, he reacted with restrictions. When he reported the following he was fully certain that his reaction was appropriate and he wanted to show that usability is something he is addressing: *"For example, I have repeatedly received the feedback: On small screens like on smartphones, you don't necessarily see the details you need to see to identify phishing emails, so I was able to recommend that you generally shouldn't open links on mobile phones if you're not sure what you're looking at."* — [Ci6] In another example, Co7 and Ci4 reported that, especially software developer would demand local administrator privileges on their machines, with both not questioning that this might be a legitimate request, but denying them with a reference to the danger that they fear comes with it.

Short Summary: A few participants were aware that usability of security mechanisms is important and can name examples, like SSO, with only one manager reporting how he implemented those concepts.

5 Discussion

In the following we discuss our findings with regard to our research questions. The majority of security managers stated that security friction was indeed a problem (especially since it can lead to negative reactions from employees that might escalate through the hierarchies). They could easily name cases where friction occurred (e. g., when they restricted access to certain programs) and some also showed understanding about concepts of usability (e. g., when they suggested that password managers or SSO would reduce the password load). However, those considerations played little to no role when they designed security policies, purchased new security products or implemented new security measures. They were able to explain friction, but could rarely name examples of how they mitigated it in their organizations – beyond suppressing friction symptoms by appeasing upset employees. While they reported knowing how to measure friction and got insights into employees demands – mainly through personal talks – they did not do so in practice, or the measurement results did not change the outcome.

Here, the lack of diversity in the security sector – also reflected in the male security leaders we recruited exclusively – becomes an obvious challenge. Within Kocksch et al.'s [57] approach to security as a discipline of care, it is assumed that such a caring approach is characterized by refraining from blaming and attributing responsibility, and instead viewing security as a collaborative and collective achievement [25]. One of our assumptions about why participants could not put usability into practice is that caring work is often feminized (and made invisible) [61] and thus is not taken into account by the male-dominated industry, by which we do not mean to promote that "women" should be declared solely responsible for security [57].

In the academic community, it was established more than a decade ago that small security demands can have a large financial impact [45, 46, 74]. Our results suggest that this knowledge still did not find its way to security (management) practice yet. However, we can not blame the security managers [77]: they are paid to make organizations secure, they have to report numbers to the business leadership that show how investments reduce security risks. The costs of security friction and their reduction is not part of those numbers, not written in norms and regulations they try to implement and

is not part of a security professional's training curriculum. Basic usability and economics concepts need to become part of that curriculum – security professionals don't have time to read research papers in usable security. In the rest of this discussion we will recapitulate some of our findings in more depth, before we derive recommendations (Section 5.1).

No Measurements = No Insights The results show that most of the interviewed security managers (CISOs and consultants) are aware of the importance of measuring friction, and that considering the other demands employees have to meet was essential for security measures to work. This recognition of human aspects of IT security contrasts other findings [77] that showed that security managers see users in a negative way. Many managers (mostly consultants) described that friction measurements were often a "one and done" solution in organizations, with no real follow-up to see long term effects. Their preferred way of measuring friction, casual talks by the coffee machine, might play into this: even though the gathering of real world experiences is recommendable, the lack of planning and structure might impair an effective, long term measurement of friction. Friction, therefore, might remain in the organization without security managers knowing about it.

Perception From what we gathered in the present study, security managers hold employees' needs and wishes in high regard, citing them as paramount for the effectiveness of security measures. Security managers, naturally, care a lot about the security in the organization, but seem to rely too much on official security regulations and guidelines. These rules are perceived, not only as a practical aid for making decisions about security, but also as an excuse if the implemented measures are not accepted. This may lead to security managers seeing themselves as more of communicators of rules instead of solution-finders. Similarly, in earlier work, CISOs have been shown to appear as "interpreters" of security [22].

Considering the view that security managers have of themselves, as communicators of rules, it is no wonder that their preferred method for mitigating friction is raising the awareness of employees. If employees are dissatisfied with security measures, regulations are brought up as a sort of "knockout argument" to mitigate non-compliance. And to mitigate friction in general, security managers want employees to know about these rules and why they need to be followed.

For the security managers, friction is something which is often seen as inevitable when implementing new security measures, and when it appears, employees are predicted to circumvent them. Negative reactions by the employees are then seen as logical and even important, even though the solution for this then seems to be reiterating the necessity of these measures.

Causes Regulatory security requirements were one of the main causes of friction according to most of the security managers. The compliance of these regulations was often seen as "above" the wishes of employees, leading to unhappiness and friction. The root cause of this security managers' view may lay in their relationship with the regulating institutions: because of a lack of time and an abundance of stress [70], managers need to, in some way or other, trust these institutions and their regulations and guidelines, assuming that a lot of thoughts must have been put into them [81]. When these are seen as perfect, internal security policies are seemingly also adapted to this strictness, leading to a constant balancing of regulation and employees' wishes, with regulation coming out as the winner. A similar case seems to apply to certifications: to get these, as seen by security managers, important certifications, employees are "walked over". This lack of consideration might be caused by a complex and expensive certification process [51], which needs a lot of resources and, in turn, prioritizes this over the employees.

What only a few security managers described as a cause was *bad IT*: security mechanisms that are just badly designed and hindering *security hygiene* – which is described as a necessary prerequisite for all further measures, such as increasing employees' understanding [47, 80]. The foregoing of useful security in favor of cheap products and mechanisms, sometimes referred to as "security debt" [73], slows down the implementation of usable security measures in the organizations. The cause of this is, possibly, a lack of resources for security in the investigated companies, which some security managers also mentioned. As the results show, this not only applies to technical factors or training- and awareness programs for the employees, but also to the measurements of friction: measurements are often done casually, or quickly, without following through in a structured way. This lack of "success"-measurement makes it impossible to get a clear view of the friction caused by security measures. And the reasoning for this, a lack of resources, is probably a dangerous misconception, which might result in a "slippery slope" into even more investments in the future: if friction measurements are neglected, inappropriate security measures can secretly pile up friction in the organization, increasing the cost of achieving compliance even further because of the need for more measurements or, worse, constant monitoring of employees [13].

Impact Our analysis revealed that participants consider friction in security as a source of risk. Although security can be increased through certain measures, the sword of Damocles can also hang over the security of the organization: in the case of the perception of friction, the employees tend to circumvent the required measures or change to a shadow security behavior [55, 56], where they try to keep the security level high, but following their own rules. This finding suggests that security managers are at least aware of the friction between

the actual work tasks and security measures. A deterioration in the quality of the relationship between security staff and employees was also described by security managers as a possible consequence of the perceived friction, which echoes the findings of Menges et al. [66]. Other negative impacts such as decreasing motivation, frustration and anger were also described. Overall, employees would feel disturbed by security in the performance of their actual work tasks and feel restricted in their freedom and productivity. These reactions of employees can have negative consequences for the economic efficiency of an organization: discouragement to carry out security-related tasks, increasing fluctuation rates, etc.

The impacts we identified are not only known in the context of security, but also in the area of safety research. For example, the challenge of *work-safety tension* has already been studied by some researchers [65, 88, 90]. As Brostoff & Sasse [17] have already made clear, there are differences between safety and security, but these two domains share, for example, the fact that they are secondary goals for employees, while they have to complete their primary tasks. Safety research has shown that when employees are in tension between work tasks and safety, they tend to prefer the productive path that requires unsafe behavior, which means that such a conflict of goals always leads to a violation of safety-related rules [16]. However, safety research, as the much older discipline, has managed to translate their findings into organizational practice. There, for example, environments have to be changed to reduce the impact on employees, and only if this is not possible the employees need to be warned or actively act. Something that did not find its way in security practice yet (see also Section 5.1).

Mitigation Our participants primarily tried to mitigate friction by raising awareness: explaining the importance of security to employees, in the hope that they would accept friction is unavoidable and stop complaining (see Section 4.5). This may work, up to a point, in cases where employees were unaware or severely underestimated a risk and the communication is convincing – but not if employees' compliance budget [13] is exhausted. Behavioral science has clearly shown that trying to increase motivation to adopt a new behavior when effort is high works only in the short-term, followed by a motivational crash [32], and the study by Poller et al. [75] documented a real-world case of a security intervention creating huge enthusiasm for secure development practice, followed by 'slipping back' into old insecure routines [80]. Instead of focusing on reducing or removing friction, security managers refer to security standards and regulation as their touchstone, which they also use to deflect employees' demands for lower-effort security. Some of our participants were aware of this being a problem and at least consider changing the rules to incorporate the needs of employees. Some participants claimed that they would like to (personally) *talk* to employees, since this can be a first step towards building a relationship [10, 66]. However, they mostly want to talk about risks and try to convince employees to just accept the friction, rather than addressing on the root causes, and adapt security to employees' needs and routines. One participant explained that he wanted employees to understand that complex passwords were necessary, and did not even consider the many usable alternatives available, such as password managers and passwordless authentication solutions [6, 79].

5.1 Recommendations for Industry

Here we derive recommendations for industry with the goal to strengthen the position of usable security in security (management) practice. Measurements of friction were seen as important by many of our participant, but they currently do not see a viable route for reducing or removing it. They seemed to consider it their job – and their job alone – to make security work. This 'lone security hero' perspective means they are afraid to ask for help [21]: the organization of resources for reducing friction, or a helping hand from colleagues running business processes and other organizational functions. And whilst they appreciated casual conversations with employees, they saw them as receivers of security knowledge and directives, not as partners in developing usable security. They need to change their perspective and build relationships, and also take a systematic approach to obtain feedback from employees, identifying friction hotspots, and engaging employees to co-design security, and engage in constantly learning. From an organizational perspective, to facilitate the communication and decision-making about friction, the currently hidden costs need to be tracked and made explicit to organizational leadership. If the possible losses [45] of neglecting security friction are made clear, decision-makers are incentivized to invest in the mitigation of it.

Usable Security Training Security managers like security certificates. Among our participants, the majority earned common certificates like CISM, CISSP or CISA (this was not only true for those managers we selected because of the criteria of having certificates, but also for those we selected because of their academic degree). Those trainings do not solely focus on technical security measures, but also on *softer* topics like secure operations, organizational risk management or secure software development. This is a perfect place to also include topics of usable security, e. g., as part of the software development life cycle [39] or in the risk-cost calculations they learn about. While efforts are made to include usable security topics into (traditionally technology-heavy) academic information security programs [35], they come too late to reach security professionals that are already in the industry for years or decades, like our participants. Certificates, that need to be renewed every few years can help to fill the usable security knowledge gap.

Usable Security Norms We find that regulations and norms prevent the implementation of usable security principles and the reduction of security friction – at least the security managers use those as an excuse for not considering such. And indeed: while ISO27001 (and its implementation guideline ISO27004) and BSI Basic Protection [36] demand password policies, security awareness training, phishing simulations, restrictions on local machines, etc., they do not consider the friction that those measures cause and the costs they raise. One could argue that the sole purpose of these norms is to set security standards, but we argue that the underlying goal is to reduce the (financial) impact of attacks on the organizations. This, however, includes a balance between risk and investments and security friction covertly raises the costs of investments – so they need to be included in these calculations. Norms are a good starting point, as previous authors already proposed for software development standards [39, 41].

Positive KPIs If usable security principles are correctly applied they reduce costs. While some positive effects are rather hard to measure (e. g., the reduction of mental workload) others are easier to measure, e. g., biometric authentication reduces the time employees have to spend on authentication tasks [71], as does SSO [53]. Technical logs, observations and surveys can be the basis for measurements. Subsequently, those can be used in KPIs and reports to the management to showcase the impact of usable security considerations and to make a clear statement for further improvement. The possible fatal security consequences of low usability, which many studies show [40, 78, 87], as well as the economic benefit from reducing the aforementioned different costs should be presented to the management in a clear way to accentuate the importance of usable security. As long as the security managers themselves do not see the necessity of such, the vendors of the products themselves should implement the measurements and report the usability advantages – they could even advertise their products through those numbers.

Security Champions Security managers rightly highlighted the need to measure friction by being "close" to employees and talking with them directly. Still, many security managers (mostly consultants) described that this was often not followed through in the organization. The implementation of the so-called *security champions* – employees who not necessarily have a background in security and who are intrinsically motivated to improve the security in their teams [1, 7, 52], and who have regular contact with the security teams – could help bridging the gap between employees and security managers, allowing this role, which would represent various employees, to be an economic contact point of friction-measurement. Security champions can also help by being "bottom-up" agents [14], who question security policies that may be too strict for employees to follow.

Learning from Safety For the implementation of feasible security measures and for mitigating security friction, organizations and security managers could use the knowledge and recommendations already gained from safety research, for example: McGonagle & Keth [65] suggest monitoring the level of tension (*friction*) in the context of safety (*security*). By this they mean the explicit monitoring of the tension perceived by employees that interferes or conflicts with the effective completion of their actual tasks. They also propose a participatory approach in which employees have the opportunity to communicate those specific aspects of their work that prevent them from doing their job safely (securely). Furthermore, they should be involved in the development of effective and efficient solutions, based on the idea that they are experts of their own work.

6 Conclusion

In this paper we reported how ($n = 14$) security managers perceive security friction in organizations. While they deem friction as a problem, they have no working strategies on how to mitigate it, rely on appeasing enraged employees, and do not consider such in their own work when creating policies and implementing measures. We conclude that security managers lack the necessary support to consider friction, namely the appropriate training, KPIs that make a case for usable security, and norms that demand it. The security industry knows and talks about usable security, and focuses on what steps they take in practice to make it happen. Identifying and dealing with security friction is an essential first step towards making security usable, but we find the managers do not identify and tackle it. For usable security research this means considering how we make our knowledge and tools more accessible to this particular user group. Our study aimed at evaluating how usable security works in organizational practice. More similar studies, especially working with those that are responsible for security – the security and business managers – are necessary to increase the impact that usable security research can have towards organizational practice. Further work can also be built around the evaluation of our suggestions for industry, e. g., positive usability KPIs in organizations. The security managers in our study all worked for rather big organizations. Further studies should also study security management in small enterprises.

Acknowledgments

We would like to thank all managers who took part in our study. Thanks to the four anonymous reviewers for their helpful feedback. Thanks to Julian Becker for his help with the literature review. Thanks to Steve Ehleringer, Maximilian Golla, Stefan Horstmann, and Konstantin Fischer for their support and proof-reading. The work was supported by

the PhD School "SecHuman – Security for Humans in Cyberspace" by the federal state of NRW, Germany and partly also by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2092 CASA – 390781972.

References

[1] Hege Aalvik. Towards an effective security champions program. Master's thesis, NTNU, 2022.

[2] Chon Abraham, Dave Chatterjee, and Ronald R. Sims. Muddling through cybersecurity: Insights from the U.S. healthcare industry. *Business Horizons*, 62(4):539–548, 2019.

[3] Anne Adams and Martina Angela Sasse. Users are not the enemy. *Communications of the ACM*, 42(12):40–46, 1999.

[4] Eirik Albrechtsen. A qualitative study of users' view on information security. *Computers & Security*, 26(4):276–289, 2007.

[5] Eirik Albrechtsen and Jan Hovden. The information security digital divide between information security managers and users. *Computers & Security*, 28(6):476–490, 2009.

[6] Nora Alkaldi and Karen Renaud. Why Do People Adopt, or Reject, Smartphone Password Managers? In Karen Renaud and Melanie Volkamer, editors, *Proceedings 1st European Workshop on Usable Security*, Reston, VA, 2016. Internet Society.

[7] Moneer Alshaikh and Blair Adamson. From awareness to influence: Toward a model for improving employees' security behaviour. *Personal and Ubiquitous Computing*, 25(5):829–841, 2021.

[8] Ashley Baines Anderson, Atif Ahmad, and Shanton Chang. Competencies of cybersecurity leaders: A review and research agenda. *ICIS 2022 Proceedings*, 2022.

[9] Ginger Armbruster, Jan Whittington, and Barbara Endicott-Popovsky. Strategic Communications Planning for a CISO: Strength in Weak Ties. In *Journal of The Colloquium for Information Systems Security Education*, volume 2, page 10, 2014.

[10] Debi Ashenden and Darren Lawrence. Security Dialogues: Building Better Relationships between Security and Business. *IEEE Security & Privacy*, 14(3):82–87, 2016.

[11] Debi Ashenden and Angela Sasse. CISOs and organisational culture: Their own worst enemy? *Computers & Security*, 39:396–405, 2013.

[12] Michael Bartsch. Woher nehmen, wenn nicht stehlen – oder wo haben Sie Ihren CISO her? (German). In Michael Bartsch and Stefanie Frey, editors, *Cybersecurity Best Practices*, pages 261–269. Springer Fachmedien Wiesbaden, Wiesbaden, 2018.

[13] Adam Beautement, M Angela Sasse, and Mike Wonham. The compliance budget: managing security behaviour in organisations. In *Proceedings of the 2008 New Security Paradigms Workshop*, pages 47–58, 2008.

[14] Ingolf Becker, Simon Parkin, and M Angela Sasse. Finding security champions in blends of organisational culture. *Proc. USEC*, 11, 2017.

[15] Jim Blythe, Ross Koppel, and Sean W Smith. Circumvention of security: Good users do bad things. *IEEE Security & Privacy*, 11(5):80–83, 2013.

[16] Sebastian Brandhorst and Annette Kluge. When the tension is rising: a simulation-based study on the effects of safety incentive programs and behavior-based safety management. *Safety*, 7(1):9, 2021.

[17] Sacha Brostoff and M Angela Sasse. Safe and sound: a safety-critical approach to security. In *Proceedings of the 2001 workshop on New security paradigms*, pages 41–50, 2001.

[18] Deanna D Caputo, Shari Lawrence Pfleeger, M Angela Sasse, Paul Ammann, Jeff Offutt, and Lin Deng. Barriers to usable security? three organizational case studies. *IEEE Security & Privacy*, 14(5):22–32, 2016.

[19] Jessica Colnago, Summer Devlin, Maggie Oates, Chelse Swoopes, Lujo Bauer, Lorrie Cranor, and Nicolas Christin. "it's not actually that horrible": Exploring adoption of two-factor authentication at a university. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–11, New York, NY, USA, 2018. Association for Computing Machinery.

[20] W Alec Cram, Jeffrey G Proudfoot, and John D'Arcy. When enough is enough: Investigating the antecedents and consequences of information security fatigue. *Information Systems Journal*, 31(4):521–549, 2021.

[21] Joseph Da Silva. Cyber security and the Leviathan. *Computers & Security*, 116:102674, 2022.

[22] Joseph Da Silva and Rikke Bjerg Jensen. 'cyber security is a dark art': The ciso as soothsayer. *arXiv preprint arXiv:2202.12755*, 2022.

[23] Darren Death. *The CISO Role within US Federal Government Contracting Organizations: A Delphi Study*. PhD thesis, Capella University, 2021.

[24] Verena Distler, Marie-Laure Zollinger, Carine Lallemand, Peter B. Roenne, Peter Y. A. Ryan, and Vincent Koenig. Security - visible, yet unseen? In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–13, New York, NY, USA, 2019. Association for Computing Machinery.

[25] Paul Dourish, Rebecca E Grinter, Jessica Delgado De La Flor, and Melissa Joseph. Security in the wild: user strategies for managing security as an everyday, practical problem. *Personal and Ubiquitous Computing*, 8:391–401, 2004.

[26] Jonathan Dutson, Danny Allen, Dennis Eggett, and Kent Seamons. Don't punish all of us: Measuring user attitudes about two-factor authentication. In *2019 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, pages 119–128, New York, 2019. IEEE.

[27] Erastus Karanja. The role of the chief information security officer in the management of IT security. *Inf. Comput. Secur.*, 25:300–329, 2017.

[28] Erastus Karanja and Mark A. Rosso. The Chief Information Security Officer: An Exploratory Study. *Journal of International Technology and Information Management*, 26:23–47, 2017.

[29] Florian M Farke, Lennart Lorenz, Theodor Schnitzler, Philipp Markert, and Markus Dürmuth. " you still use the password after all"—exploring fido2 security keys in a small company. In *Proceedings of the Sixteenth USENIX Conference on Usable Privacy and Security*, pages 19–35, 2020.

[30] Todd Fitzgerald and Micki Krause. *CISO leadership: Essential principles for success*. CRC Press, 2007.

[31] Flynn Wolf, Adam J. Aviv, and Ravi Kuber. Security Obstacles and Motivations for Small Businesses from a CISO's Perspective. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 1199–1216. USENIX Association, 2021.

[32] Brian J Fogg. *Tiny habits: The small changes that change everything*. Eamon Dolan Books, 2019.

[33] Steven Furnell. Security fatigue. *Encyclopedia of Cryptography, Security and Privacy*, pages 1–5, 2019.

[34] Simson Garfinkel and Heather Richter Lipford. Usable security: History, themes, and challenges. *Synthesis Lectures on Information Security, Privacy, and Trust*, 5(2):1–124, 2014.

[35] Binto George, Martha Klems, and Anna Valeva. A method for incorporating usable security into computer security courses. In *Proceeding of the 44th ACM Technical Symposium on Computer Science Education*, SIGCSE '13, page 681–686, New York, NY, USA, 2013. Association for Computing Machinery.

[36] German Federal Office for Information Security. IT-Grundschutz-Compendium. Standard, BSI – German Federal Office for Information Security, Bonn, DE, 2022.

[37] Brain Glass, Graeme Jenkinson, Yuqi Liu, M Angela Sasse, and Frank Stajano. The usability canary in the security coal mine: A cognitive framework for evaluation and design of usable authentication solutions. *arXiv preprint arXiv:1607.03417*, 2016.

[38] Marilu Goodyear, Holly T. Goerdel, Shannon Portillo, and Linda Williams. Cybersecurity Management In the States: The Emerging Role of Chief Information Security Officers. *SSRN Electronic Journal*, 2010.

[39] Marco Gutfleisch, Jan H. Klemmer, Niklas Busch, Yasemin Acar, M. Angela Sasse, and Sascha Fahl. How does usable security (not) end up in software products? results from a qualitative interview study. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 893–910, New York, 2022. IEEE.

[40] Marco Gutfleisch, Maximilian Peiffer, Selim Erk, and Martina Angela Sasse. Microsoft office macro warnings: A design comedy of errors with tragic security consequences. In *Proceedings of the 2021 European Symposium on Usable Security*, pages 9–22, 2021.

[41] Marco Gutfleisch, Markus Schöps, Jonas Hielscher, Mary Cheney, Sibel Sayin, Nathalie Schuhmacher, Ali Mohamad, and M. Angela Sasse. Caring about iot-security – an interview study in the healthcare sector. In *Proceedings of the 2022 European Symposium on Usable Security*, EuroUSEC '22, page 202–215, New York, NY, USA, 2022. Association for Computing Machinery.

[42] Julie M. Haney and Wayne G. Lutters. The work of cybersecurity advocates. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '17, page 1663–1670, New York, NY, USA, 2017. Association for Computing Machinery.

[43] Julie M Haney and Wayne G Lutters. " it's scary... it's confusing... it's dull": How cybersecurity advocates overcome negative perceptions of security. In *SOUPS@ USENIX Security Symposium*, pages 411–425, Berkeley, 2018. USENIX.

[44] Julie M. Haney and Wayne G. Lutters. Cybersecurity Advocates: Discovering the Characteristics and Skills of an Emergent Role. *Information and Computer Security*, 29(3), 2021.

[45] Cormac Herley. So long, and no thanks for the externalities: the rational rejection of security advice by users. In *Proceedings of the 2009 workshop on New security paradigms workshop*, pages 133–144, 2009.

[46] Cormac Herley. More is not the answer. *IEEE Security & Privacy*, 12(1):14–19, 2013.

[47] Jonas Hielscher, Annette Kluge, Uta Menges, and M. Angela Sasse. "taking out the trash": Why security behavior change requires intentional forgetting. In *New Security Paradigms Workshop*, NSPW '21, page 108–122, New York, NY, USA, 2022. Association for Computing Machinery.

[48] Jonas Hielscher, Uta Menges, Simon Parkin, Annette Kluge, and M. Angela Sasse. "Employees Who Don't Accept the Time Security Takes Are Not Aware Enough": The CISO View of Human-Centred Security. In *32st USENIX Security Symposium (USENIX Security 23)*, Boston, MA, August 2023. USENIX Association.

[49] Val Hooper and Jeremy McKissack. The emerging role of the CISO. *Business Horizons*, 59(6):585–591, 2016.

[50] Nicolas Huaman, Bennet von Skarczinski, Christian Stransky, Dominik Wermke, Yasemin Acar, Arne Dreißigacker, and Sascha Fahl. A Large-Scale interview study on information security in and attacks against small and medium-sized enterprises. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 1235–1252. USENIX Association, August 2021.

[51] Mike Hulshof and Maya Daneva. Benefits and challenges in information security certification–a systematic literature review. In *Business Modeling and Software Design: 11th International Symposium, BMSD 2021, Sofia, Bulgaria, July 5–7, 2021, Proceedings 11*, pages 154–169. Springer, 2021.

[52] Martin Gilje Jaatun and Daniela Soares Cruzes. Care and feeding of your security champion. In *2021 International Conference on Cyber Situational Awareness, Data Analytics and Assessment (CyberSA)*, pages 1–7, New York, 2021. IEEE, IEEE.

[53] Neely James, Shruti Marwaha, Stacie Brough, and Thomas T John. Impact of single sign-on adoption in an assessment triage unit: A hospital's journey to higher efficiency. *JONA: The Journal of Nursing Administration*, 50(3):159–164, 2020.

[54] Julia H Allen, Gregory Crabb, Pamela Curtis, Brendan Fitzpatrick, Nader Mehravari, and David Tobar. Structuring the Chief Information Security Officer Organization.

[55] Iacovos Kirlappos, Simon Parkin, and M. Angela Sasse. "Shadow security" as a tool for the learning organization. *ACM SIGCAS Computers and Society*, 45(1):29–37, 2015.

[56] Iacovos Kirlappos, Simon Parkin, and M. Angela Sasse. Learning from "Shadow Security": Why Understanding Non-Compliant Behaviors Provides the Basis for Effective Security. In Matthew Smith and David Wagner, editors, *Proceedings 2014 Workshop on Usable Security*, Reston, VA, February 23, 2014. Internet Society.

[57] Laura Kocksch, Matthias Korn, Andreas Poller, and Susann Wagenknecht. Caring for IT Security. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–20, 2018.

[58] Udo Kuckartz. *Qualitative inhaltsanalyse (German)*. Beltz Juventa, 2012.

[59] Benedikt Lebek, Jörg Uffen, Michael H. Breitner, Markus Neumann, and Bernd Hohler. Employees' information security awareness and behavior: A literature review. In *2013 46th Hawaii International Conference on System Sciences*, pages 2978–2987, New York, 2013. IEEE.

[60] Michelle R. Lowry, Anthony Vance, and Marshall D. Vance. Inexpert Supervision: Field Evidence on Boards' Oversight of Cybersecurity. *SANS*, 2021.

[61] Aryn Martin, Natasha Myers, and Ana Viseu. The politics of care in technoscience. *Social studies of science*, 45(5):625–641, 2015.

[62] Peter Mayer, Nina Gerber, Ronja McDermott, Melanie Volkamer, and Joachim Vogt. Productivity vs security: mitigating conflicting goals in organizations. *Information & Computer Security*, 2017.

[63] Peter Mayer, Collins W. Munyendo, Michelle L. Mazurek, and Adam J. Aviv. Why users (don't) use password managers at a large educational institution. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 1849–1866, Boston, MA, August 2022. USENIX Association.

[64] Sean B. Maynard, Mazino Onibere, and Atif Ahmad. Defining the Strategic Role of the Chief Information Security Officer. *Pacific Asia Journal of the Association for Information Systems*, pages 61–86, 2018.

[65] Alyssa K McGonagle and Lisa M Kath. Work-safety tension, perceived risk, and worker injuries: A meso-mediational model. *Journal of safety research*, 41(6):475–479, 2010.

[66] Uta Menges, Jonas Hielscher, Annalina Buckmann, Annette Kluge, M. Angela Sasse, and Imogen Verret. Why IT Security Needs Therapy. In *Computer Security. ESORICS 2021 International Workshops*, volume 13106 of *Lecture Notes in Computer Science*, pages 335–356. Springer International Publishing, Cham, 2022.

[67] Eric Molin, Kirsten Meeuwisse, Wolter Pieters, and Caspar Chorus. Secure or usable computers? Revealing employees' perceptions and trade-offs by means of a discrete choice experiment. *Computers & Security*, 77:65–78, 2018.

[68] Tyler Moore, Scott Dynes, and Frederick R. Chang. Identifying how firms manage cybersecurity investment. *Available: Southern Methodist University.*, 32, 2015.

[69] Tabisa Ncubukezi. Human errors: A cybersecurity concern and the weakest link to small businesses. In *Proceedings of the 17th International Conference on Information Warfare and Security*, page 395, 2022.

[70] Calvin Nobles. Stress, Burnout, and Security Fatigue in Cybersecurity: A Human Factors Problem. *HOLISTICA–Journal of Business and Public Administration*, 13(1):49–72, 2022.

[71] MO Oloyede, AO Adedoyin, and KS Adewole. Fingerprint biometric authentication for enhancing staff attendance system. *International Journal of Applied Information Systems*, 2013.

[72] Jon Oltsik, Candy Alexander, and CISSP CISM. The life and times of cybersecurity professionals. *ESG and ISSA: Research Report*, 2017.

[73] Simon Parkin, Simon Arnell, and Jeremy Ward. Change that respects business expertise: Stories as prompts for a conversation about organisation security. In *New Security Paradigms Workshop*, NSPW '21, page 28–42, New York, NY, USA, 2021. Association for Computing Machinery.

[74] Simon Parkin, Aad van Moorsel, Philip Inglesant, and M. Angela Sasse. A stealth approach to usable security: Helping it security managers to identify workable security solutions. In *Proceedings of the 2010 New Security Paradigms Workshop*, NSPW '10, page 33–50, New York, NY, USA, 2010. Association for Computing Machinery.

[75] Andreas Poller, Laura Kocksch, Sven Türpe, Felix Anand Epp, and Katharina Kinder-Kurlanda. Can security become a routine? a study of organizational change in an agile software development group. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, pages 2489–2503, 2017.

[76] Gerald V Post and Albert Kagan. Evaluating information security tradeoffs: Restricting access can interfere with user tasks. *Computers & Security*, 26(3):229–237, 2007.

[77] Lena Reinfelder, Robert Landwirth, and Zinaida Benenson. Security Managers Are Not The Enemy Either. In Stephen A. Brewster, Geraldine Fitzpatrick, Anna L. Cox, and Vassilis Kostakos, editors, *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*, pages 1–7, New York, New York, USA, 2019. ACM Press.

[78] Scott Ruoti, Jeff Andersen, Daniel Zappala, and Kent Seamons. Why johnny still, still can't encrypt: Evaluating the usability of a modern pgp client. *arXiv preprint arXiv:1510.08555*, 2015.

[79] Scott Ruoti, Brent Roberts, and Kent Seamons. Authentication melee: A usability analysis of seven web authentication systems. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, page 916–926, Republic and Canton of Geneva, CHE, 2015. International World Wide Web Conferences Steering Committee.

[80] Angela Sasse, Jonas Hielscher, Jennifer Friedauer, and Annalina Buckmann. Rebooting it security awareness – how organisations can encourage and sustain secure behaviours. In *European Symposium on Research in Computer Security*, pages 1–18, Berlin, 09 2022. Springer, Springer.

[81] M Angela Sasse, Matthew Smith, Cormac Herley, Heather Lipford, and Kami Vaniea. Debunking security-usability tradeoff myths. *IEEE Security & Privacy*, 14(5):33–39, 2016.

[82] Martina Angela Sasse, Sacha Brostoff, and Dirk Weirich. Transforming the 'weakest link'—a human/computer interaction approach to usable and effective security. *BT technology journal*, 19(3):122–131, 2001.

[83] Conrad Shayo and Frank Lin. An exploration of the evolving reporting organizational structure for the chief information security officer (ciso) function. *Journal of Computer Science*, 7(1):1–20, 2019.

[84] Steve Sheng, Levi Broderick, Colleen Alison Koranda, and Jeremy J Hyland. Why johnny still can't encrypt: evaluating the usability of email encryption software. In *Symposium on usable privacy and security*, pages 3–4. ACM, 2006.

[85] Ben Shneiderman, Catherine Plaisant, Maxine S Cohen, Steven Jacobs, Niklas Elmqvist, and Nicholas Diakopoulos. *Designing the user interface: strategies for effective human-computer interaction*. Pearson, 2016.

[86] Brian Stanton, Mary F Theofanos, Sandra Spickard Prettyman, and Susanne Furman. Security fatigue. *It Professional*, 18(5):26–32, 2016.

[87] Christian Stransky, Oliver Wiese, Volker Roth, Yasemin Acar, and Sascha Fahl. 27 years and 81 million opportunities later: Investigating the use of email encryption for an entire university. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 860–875. IEEE, 2022.

[88] Chris B Stride, Nick Turner, M Sandy Hershcovis, Tara C Reich, Chris W Clegg, and Philippa Murphy. Negative safety events as correlates of work-safety tension. *Safety science*, 53:45–50, 2013.

[89] U.S. Department of Homeland Security. The Menlo Report: Ethical Principles Guiding Information and Communication Technology Research, August 2012. https://www.caida.org/publications/papers/2012/menlo_report_actual_formatted/, as of June 2, 2023.

[90] Benjamin M Walsh, Alyssa K McGonagle, Timothy Bauerle, and Tarya Bardwell. Safety stressors: Deviant reactions to work-safety tension. *Occupational Health Science*, 4:63–81, 2020.

[91] Jake Weidman and Jens Grossklags. I like it, but i hate it: Employee perceptions towards an institutional transition to byod second-factor authentication. In *Proceedings of the 33rd Annual Computer Security Applications Conference*, ACSAC '17, page 212–224, New York, NY, USA, 2017. Association for Computing Machinery.

[92] Alma Whitten and J Doug Tygar. Why johnny can't encrypt: A usability evaluation of pgp 5.0. In *USENIX security symposium*, volume 348, pages 169–184, 1999.

[93] Dwayne Whitten. The chief information security officer: An analysis of the skills required for success. *Journal of Computer Information Systems*, 48(3):15–19, 2008.

A Interview Guide

The following interview guide is the one developed for the CISOs. The interview guide for consultants differs in how the questions are asked for *the organizations they advise*, rather than their own organizations.

1. **Looking back on your career, how and when did you first come into contact with cybersecurity?**
   (a) What were the most important turning points in your professional career so far (describe them briefly)?
   (b) What experience or training do you currently have in cybersecurity?
   (c) In what time and based on what training and/or experience were you able to acquire the greatest part of your knowledge in the field of cybersecurity?
2. **What are currently the biggest structural and/or organizational challenges you are facing in the context of cybersecurity?**
   (a) Describe briefly and compactly the form of organization in which you are embedded?
   (b) Which organizational interfaces are currently causing you the most challenges?
   (c) How do you determine whether and how your environment (management, board of directors, etc.) is sufficiently 'aware' of cybersecurity issues?
   (d) On the basis of which circumstances is the budget for cybersecurity decided?
   (e) Which activities take up most of your attention?
3. **The application of security measures (technical, organizational or administrative) usually leads to an increase in the security maturity of an organization. What effects can such measures have on employees in your organization?**
   (a) Do you try to get feedback from employees after applying new measures?
   (b) Can you give some examples of this?
   (c) How are security measures generally perceived by your employees?
4. **In everyday work, there are often business interests versus security interests to be balanced. When developing and applying security measures, do you think about the concrete effects (individual consequences) on the individual employees affected?**
   (a) How do you decide whether business interests or employee interests take precedence over security interests?
   (b) Based on which metrics, methods or techniques do you balance the respective interests?

(c) In your opinion, which personal interests of employees are legitimate and must be taken into account? Which are not?

(d) Do you know of any examples where cybersecurity measures have been adapted or even improved based on feedback from employees (briefly describe them)?

5. **It can happen that employees do not dare to show negative reactions or do not know who to turn to with their criticism. What negative experiences do you think these employees would report?**

   (a) How would you describe the relationship between security and restriction (Can there be security without 'sacrifice'?)?

   (b) What do you do to keep the corresponding negative reactions as low as possible?

   (c) How do you react to negative reactions?

   (d) In your view, are such negative reactions legitimate?

6. **Can an accumulation of negative staff reactions become a problem for organizations?**

   (a) What can be the consequences for an organization if these negative reactions are not taken ito account?

   (b) Have you ever had to deal with staff reactions that were escalated (carried over the reporting line)?

   (c) What could be the triggers for such escalations?

7. **How do you find out whether the security measures can be implemented during the employees' work routines or not (process adaptability, etc.)?**

   (a) To what extent do you make an effort to understand employees' work routines?

   (b) How do you ensure that security measures can be applied by employees?

8. **How do you engage with employees' work routines?**

   (a) Which means and methods do you prefer to understand the work routines of your employees?

   (b) How important do you rate the issue on scale from 1 (not important) to 10 (very important) that security measures can be integrated into the work routines of employees?

B Accumulated Demographic Data

Table 2: Accumulated demographic data of our participants.

Gender			#	%
Male			14	*100%*
Highest (Security) Education				
Master/ Diploma Computer Science			3	*21%*
Master/ Diploma (Information) Security			3	*21%*
Vocational Training			1	*7%*
Security Certificates			4	*30%*
Other Master/ Diploma			3	*21%*
Experience in Security				
Max	38	*Average*	19.8	
Min	8	*Median*	20	
Sum	258			

C Code Book

Table 3: The code book (1/2). *Occ.*: the number of occurrences of the code in all documents.

Code	Description	Example Quote	Occ.
(Perceived) Causes of Friction	Where does friction come from? The hard hand of the participant, from norms, from management, from security itself.	*If poorly designed security measures or poorly designed awareness campaigns always lead to resistance at the beginning.*	15
ISO Norms and Regulations	All the audit and norm problems that the participants report with regards to their security strategy.	*On the one hand, we have clear regulatory requirements. That means we have to implement them and can have relatively little consideration for the people themselves.*	22
Additional Security	The participant explains that his organization needs additional security measures that the employees need to implement or follow that cause or might cause friction.	*And the bad guys are the security people, because they now demand something that wasn't necessary before, and that leads to these backlashes.*	7
Impact of Friction	(Negative) consequences of security friction on all stakeholders and the organization itself. This includes all types of negative reactions of employees and others.	*Or, even worse, is hidden. If the security measure is deactivated without those responsible realizing it.*	11
Negative Reactions	The employees dislike the friction caused by security/ they actively react negative.	*Unsightly case: An Employee, IT manager, who showed up at the workplace with a shotgun. This is a kind of escalation. Not in the way, probably, that you expected now. [...] It had to do with the fact that freedoms, in quotation marks, were restricted by standardization and harmonization.*	23
Primary vs. Secondary Task Conflict	The security task clashes with the primary task of the employees.	*And at first glance, a longer password or multiple passwords can look very banal, because, okay, then you enter one more password. But that can prevent an employee from doing the work at all, or perhaps from doing it less often or less regularly. Or he gets upset, needs additional breaks. That all has an impact.*	17
Economic Impact	Impact on the productivity, revenue, etc. of an organization through security.	*Because every organization has to generate output somehow. Or let's take the private sector: organizations have to generate revenue. And any aspect that has even the slightest negative impact on an employee's daily business, let's say, instead of single sign-on, the password has to be entered every time. That reduces the output that the employee can provide.*	10
Shadow Security/ Circumventing Security	Friction will cause circumventing security policies and measures.	*If no platform is offered for secure data exchange, then an employee has a legitimate interest. And it is precisely then that he will turn to any private means he knows, e-mail or any other cloud services, Dropbox, etc., simply for lack of an alternative that is not available.*	11
(Perceived) Solutions for Friction	The participant explains how security friction can or should be reduced in his opinion (this includes hard measures like taking security tasks away, but also soft measures like "just explaining friction to employees so that they understand and accept it").	*You simply have to find a sensible balance between what you allow and what you ban, because you can't ban everything. Instead, you have to weigh up how bad this is, what I am supposed to judge? And do I have to ban it or not?*	10

Table 4: The code book (2/2). *Occ.*: the number of occurrences of the code in all documents.

Code	Description	Example Quote	Occ.
Awareness and Communication Will Solve Friction	Awareness and/or communication with the employees will lead to an understanding of security friction.	*Security is not, so is always perceived as somewhere, yes, maybe an obstacle or something. So we're aware of that, and we try to maintain a positive image, i.e. that people can approach us at any time. But security has priority, of course.*	39
Develop Security Together	Employees and business units can co-define how security should work.	*As a matter of principle, we try to develop the solutions together with IT and pick up the teams from the business side early on.*	9
Change Security to Reduce Friction	Security is reduced or changed in order to reduce the friction experienced by the employees.	*and it may well be that there is a legitimate interest, and then we can also reconsider the solution.*	10
Measurements and Observations	Measurements that the participants or other stakeholders take to understand, learn or quantify security friction (including usable security, time effort) and working routines (that might be affected by security measures) of employees.	-	-
Methods	With which methods is the security friction measured?	*I ask questions. I go and ask people, "How's that working out for you now?"*	55
No Measurements	Security friction is not measured.	*I hardly ever observe that, that feedback is collected. No, I hardly ever observe that.*	9
Active Communication	Actively talking about security measures and friction with the employees.	*We also like it when comments come in unfiltered. That's what I said at the beginning, because it's very hierarchical, you sometimes don't feel the pulse of the employees.*	3
Usable Security	Definitions, status quo descriptions, technical measures, attitudes about usable security. E.g., to say that "security needs to be user friendly", or that "security is not compatible with usability", or that "longer passwords are unusable".	*And basically the issue of single sign-on. If I want to or have to log on to different platforms because I need different tools, different applications, different services, but I can largely cover this with single sign-on, that's an increased security feature. But at the same time, it also improves user-friendliness.*	34
Invisible Security	Security is good if it is not visible to the employees.	*So in the best case, not at all. So if we, let's stay with the example of user authentication et cetera, if that's possible, if that goes by very gently, so that we don't virtually burden the entire security with the, the employee, but rather check that quasi in the background.*	4
Hard Hand/Restrictions	The participant restricts (or wants to restrict) what employees can do and/ or pushes for a law-and-order policy.	*That's why I have to clarify restrictions somehow, you are not allowed to attach this file to an e-mail, whether you understand it or not, that's just the way it is. And if I specify something like that, then I have to think about exactly when I specify it, how I can either technically enforce it so that it is not possible. Or, on the other hand, how can I monitor people who don't comply so that I can draw their attention to it or, in the worst case, impose sanctions on them?*	9

What can central bank digital currency designers learn from asking potential users ?

Svetlana Abramova [*]
Universität Innsbruck

Rainer Böhme [*]
Universität Innsbruck

Helmut Elsinger [†]
Oesterreichische Nationalbank

Helmut Stix [†]
Oesterreichische Nationalbank

Martin Summer [†]
Oesterreichische Nationalbank

Abstract

The ongoing initiatives to offer central bank money to consumers in the form of retail central bank digital currency (CBDC) have triggered discussions on its optimal design. So far, the perspective of potential users has not been considered widely. To strengthen this, we survey 2006 Austrian residents using a tailored questionnaire on attitudes towards a digital euro, selected technical features as well as potential security and privacy concerns. Only about half of the surveyed respondents express at least some interest in a digital euro. This subsample tends to attribute more importance to security aspects than to transaction data privacy. Similarly, offline functionality is preferred over a feature to make direct payments between persons. Our findings suggest central banks to embrace a more user-centric design of CBDC. This effort should include communicating the key concepts and benefits to the potential users.

1 Introduction

The question on whether and how central banks should issue central bank money in digital form directly to consumers is high on the policy agenda. Reports and academic papers have contributed to the discussion of retail central bank digital currencies (CBDC) from various angles, including monetary policy [3, 13, 16], impact on the financial system [4, 28], and technology [5, 26]. Comparatively fewer studies have taken the perspective of potential users, let alone have applied methods to systematically collect data on a representative basis [11, 33]. To address this gap, this paper draws on a dataset collected from Austrian residents, who were asked about their interest in a digital euro. The respondents also stated their preferences on such key features of a retail CBDC as the access model, offline functionality, and person-to-person payments. They further reported the perceived importance of technical attributes, such as payments security and privacy (i. e., data protection). A series of logistic regression models is estimated and discussed with a view on informing the ongoing policy debate.

A distinctive feature of our study is that we do not only control for socio-economic factors, but also identify a typology of consumers based on their current use of payment instruments, the degree of technology-savviness, and the reported ownership of cryptocurrencies. We conjecture that these factors play distinct roles in the adoption path of a prospective digital euro. For example, users of non-cash payment instruments, tech-savvy persons, and owners of cryptocurrencies are likely to be among the first adopters of CBDC. Cryptocurrency owners deserve special attention as they have already collected experience with elements of new forms of (arguably) digital money, such as wallets or the handling of cryptographic keys. Their opinion may be more informed given that future CBDC is an abstract concept to most respondents in population surveys.

On the other hand, cash use is still widespread in many European countries. In prior work [8, 36], cash-affine users (i. e., those who prefer to pay with cash for their purchases) were found to have rather different attitudes towards payment instruments than users exercising a more flexible choice. Studying the views of cash-affine users is informative to gauge the initial "market potential" of CBDC. Their adop-

[*]Contact {svetlana.abramova | rainer.boehme} @ uibk.ac.at

[†]The views expressed here are those of the authors and do not necessarily represent the views of the Oesterreichische Nationalbank or the Eurosystem.

tion behavior might be pivotal to determine whether CBDC will develop to become a substitute or remain a complement to the existing payment instruments.

Our results show that consumers in Austria are largely unaware of a digital euro and express little interest in it when prompted. Using a series of tailored questions to elicit the preference between an account model for CBDC (inspired by online banking and card payments) and an access model using digital tokens (inspired by cryptocurrencies), we find overwhelming support for the account-based access. This result is corroborated by our findings on consumers' attitudes towards security and privacy. While the majority assigns high importance to security against fraud and theft, two attributes concerning transaction data privacy rank lowest in a list of nine general attributes: less than one third of the respondents considers it very important that individual transactions are untraceable.

Our paper makes a number of contributions. Drawing upon systematically collected data, we shed light on consumers' interest in a CBDC and their preferences regarding its key technical features. Given the innovative nature of this technology, we suggest a typology of consumer types, which aids to refine heterogeneous opinions and identify groups of prospective early adopters. Finally, we offer guidance for central banks, policy makers, and researchers that facilitates a more user-centric and empirically founded approach to CBDC design. As a high-level lesson, CBDC designers must not underestimate how exotic the concept of a digital euro is for large parts of its intended user base.

This paper is organized as follows. The next section recalls the background of this study and relates it to prior work. Section 3 describes our method, Section 4 presents the empirical results in detail, whereas Section 5 discusses the implications on a higher level. The paper closes with a brief conclusion.

2 Background

This section sets the scene. Subsection 2.1 briefly recalls the justifications for central banks' CBDC projects and relates them to the perspective of consumers studied in the present work. Subsection 2.2 introduces selected challenges in CBDC design and the associated terminology. Subsection 2.3 presents a review of closely related work. Readers familiar with these topics can safely skip this section.

2.1 Why CBDC?

According to the Bank for International Settlements (BIS), more than 75 central banks around the world are examining whether they should offer central bank money to the public not only as banknotes and coins but also in digital form [7]. This new form of money is referred to as retail CBDC.

Most central banks, including the European Central Bank (ECB), view their work on retail CBDC as a strategic project. It should enable universal access to central bank money in a future in which digital payments are becoming more important, while the payments market could be dominated by new private intermediaries, including the big global platform firms of the internet economy. While central banks' projects are in different stages of development, the majority of them are driven by administrative prudence and strategic foresight rather than the desire to phase out existing forms of money such as cash [10, 17]. Issues like the continued universal access to central bank money, control over monetary policy as well as sovereignty issues take a lot of room in the discussions of central banks and policy makers.

Consumers, by contrast, seem often unaware of these debates and currently do not exert much active pressure on central banks to offer new forms of money and payment instruments. However, a new form of digital money cannot be developed and implemented by a central bank decision alone. It needs to be adopted by users and provide functions that cater to real user needs and preferences. In our study, we want to better understand the current user perspective in order to inform the debate on CBDC.

2.2 Key CBDC design decisions

The design space for retail CBDC is large. It spans technical as well as economic and legal aspects. Our survey touches on a number of technical design decisions to be made before the launch of a CBDC that are costly (if not infeasible) to revert later. The selection of aspects was guided, on the one hand, by their relevance in the policy debate and, on the other hand, by what consumers can meaningfully state in a survey about an imagined form of money.

Account or token-based access The way how end users can access CBDC has far-reaching implications ranging from usability, privacy, security against theft and losses, perhaps including the mental model

future consumers form about money. While the technical design space is rich and not fully explored, it is commonly simplified to a dichotomy between account versus token-based access [5]. The former follows a conventional account model used in banking systems: ownership of and control over digital money is established by verifying the identity of an account holder. By contrast, the token-based model seeks to mimic the nature of banknotes and coins in digital form. Inspired by how cryptocurrencies manage access, token-based access conditions control (and hence ownership) on the mere knowledge of a secret, typically a private cryptographic key. Strictly speaking, token-based access refers to digital tokens; the model should not be confused with physical tokens (e. g., pieces of hardware) that can change hands just like cash. To illustrate the differences between account and token-based access, consider the protection against financial losses and privacy risks. The token-based model can offer more privacy by de-linking one's identity from transactions, however suffers from a higher risk of losing funds in case of stolen or forgotten keys. The loss of cryptographic keys would resemble the loss of a printed financial bearer instrument.

Offline and person-to-person payments Most consumers have experience with several of the existing electronic payment options offered by the private sector. Retail CBDC differs in the institutional arrangement and requires a new legal framework to ensure the stability of the currency in times of crises or when the demand for cash vanishes. However, it may be difficult for individuals to appreciate these social advantages in normal times and while cash is still widely used. Therefore, in order to increase the individual benefits of CBDC, policy makers may explore the idea of equipping CBDC with features that most existing electronic payments do not offer.

The features considered in our study are offline functionality and person-to-person payments. The former refers to the ability to make payments when there is no network coverage, for example in remote areas or during a temporary blackout. The latter refers to a simple way of passing money directly between individuals (i. e., without a merchant), typically in an interpersonal exchange. Scenarios include pocket money to children, donations and tips to unknown people, splitting bills, or yard sales.

Security and privacy Security and privacy are relevant non-functional properties of any payment system that processes large values or is widely adopted. As such, they set crucial boundary conditions for the design of digital currencies [25]. From the central bank's perspective, each property is costly to engineer, and certain security and privacy features are technically incompatible with each other [6].

Security primarily means that nobody except the legitimate owner can spend funds. As it is widely acknowledged that absolute security is infeasible, a broader notion of CBDC security should include the ease of becoming a proficient user, who makes few mistakes and does not fall for fraudulent requests (e. g., like phishing attempts, which cause a main security risk in online banking). The broadest notion of security from a consumer's point of view incorporates means to recover from failure, e. g., to dispute a transaction and revert payments in justified cases.

While security protects the user from unintended transactions, privacy means that intended transactions do not reveal unintended information about the transaction and the involved parties. As digital technology has matured to a level where storage of information is extremely cheap, many systems are designed to never forget. Such designs pose significant privacy risks. Electronic payments data is considered particularly sensitive as it may reveal information about individuals' wealth, attitudes, preferences, and behaviors. To protect individuals from undesirable consequences of secondary use (or misuse) of personal data that was initially collected for the purpose of payment processing, CBDCs could employ advanced technologies, some of which are still under ongoing research. These technologies support the principle of data minimization, which is adopted in many data protection laws, chiefly the EU's General Data Protection Regulation.

However, the deliberate choice to offer privacy is also subject to policy discussion: should CBDC offer the same level of anonymity and untraceability as cash payments, or should some data be retained and certain secondary uses be enabled? For example, law enforcement agencies could be allowed in justified cases to "follow the money" in order to solve crimes. This promises an increase in security at the cost of privacy. Such trade-offs appear in many forms. For example, having a record about a payee's identity makes it easier (if not enables) for the payer to claim back misdirected payments through the legal system.

2.3 Related work

The literature on CBDC has grown quickly recently. Most of the papers in policy discussions are concerned with strategic considerations as well as technological and economic analyses [7].

Research on user expectations, their preferences for digital central bank money, and the perspective on security and privacy aspects has remained relatively scarce. We are aware of four related empirical studies, two of which are based on surveys [11, 33], one involves focus groups [27], and one uses a mixed-methods approach [31]. An alternative way is to use survey data on existing payment instruments in order to predict demand for CBDC with structural models [24, 29].

The OMFIF study [33] analyzes survey data from more than 13,000 individuals in the age range from 16 to 75. The respondents were recruited from an online panel covering 12 countries. The survey focused on trust in different institutions as potential issuers of digital money, on the importance of different characteristics of payment methods from the user perspective, as well as the subjective assessment of some properties of different payment methods, such as speed, safety etc. The study finds that an openness to the prospective adoption of digital money rises with income and education but declines with age. Safety from theft and fraud ranks highest in the preferred ideal characteristics of a payment method.

An early survey study on the potential adoption of CBDC was [11]. The paper analyzes a sample of 3,293 individuals recruited from an online panel of Dutch residents aged 16 and above. In line with the OMFIF survey, the authors find that potential early adopters of a CBDC are younger, higher educated, and earn higher incomes. While the majority of respondents have never heard of CBDC before participating in the survey, when prompted about 50% expressed a general interest in CBDC, both as a means of payment and as a savings instrument.

The most recent study on consumer attitudes and expectations of a digital currency was published in a report by Kantar Public [27], which documents results from various focus groups analyzed for countries in the euro area on behalf of the ECB. These results are hence not based on representative surveys. Like in [33, 11], few people, including individuals who are characterized as "tech-savvy," have heard about a digital euro. The respondents would value universal access, ease and simplicity of use as well as speed and security most highly as properties of digital money in general. While people in the Kantar study do rank security highly, they do not express very strong concerns regarding privacy of transaction data.

The report by Maiden Labs [31] used both qualitative interviews and a national survey of 1,319 US citizens to learn about their relationship to and use of payment systems. In contrast to the other works, the surveyed respondents were found to be concerned about financial privacy risks, in particular with respect to their own social circles.

User experience in the domain of cryptocurrencies is another research area peripheral to our work. Empirical studies have shown that cryptocurrency owners have inadequate mental models of decentralized systems and crypto wallets serving as payment gateways [21, 30, 32, 38]. These tools, many of which were originally designed with little to no usability in mind, are often perceived to be complex and prone to security and privacy pitfalls. With CBDC initiatives being still in a formative stage, our work strives to advocate for integrating user perspectives into the design process at early stages.

Compared to this state of the art, this paper and its accompanying technical report[1] offer insights from representative—to the extent possible in times of a pandemic—data of a country in the euro area (Austria, 9 million residents, € 50,000 GDP per capita). A new breakdown of results by consumer types helps us to map the heterogeneity in attitudes and user needs with regard to payments in general, and possible future use of CBDC in particular. Collecting data in a country with a relatively high share of cryptocurrency ownership, and at the same time a large sub-group of users who have a strong preference for cash, allows us to contrast the needs and expectations of potential early adopters better than looking at broad mean values.

3 Method

This section documents the data collection, defines consumer types and other control variables, and explains the specification of the regression analyses.

3.1 Data

Our data are collected as part of a survey commissioned by the Austrian Central Bank ("OeNB

[1] https://www.oenb.at/dam/jcr:e3199ed9-0b24-4df5-aac9-52c12c2fbe72/WP_241.pdf

Barometer 2021/1"). The survey is undertaken semi-annually and mainly focuses on economic sentiments and expectations. The questionnaire used in this paper has been devised by the authors and appended as a special module to the regular survey.[2] After several iterations of pretests, it was administered between 18 June and 20 July 2021 by the Austrian IFES institute. The sample consists of 2,006 Austrian residents from age 16 and above, sampled at random from a database of phone numbers. The sampled persons were asked whether they would like to participate in the survey via telephone interview (CATI, 353 interviews) or online interview (CAWI, 1653 interviews). This mixed-mode design differed from past OeNB Barometer surveys, which were based on in-person interviews only. This choice had to be made due to the pandemic situation.

3.2 Approach

To analyze individuals' attitudes towards CBDC, we estimate regression models which evaluate the effect of different socio-economic characteristics. In addition, we consider three types of consumers: cryptocurrency owners, tech-savvy persons, and cash-affine consumers.

Consumer types Our typology of consumers is based on the following considerations. First, since both future CBDCs and cryptocurrencies represent some form of "digital money," we assume that it is easier for cryptocurrency owners to imagine handling a digital euro and the necessary elements (e. g., wallets, cryptographic keys). Therefore, cryptocurrency owners may serve as valuable informants to the designers of CBDCs concerning technical aspects and user experience. In addition, collecting individuals' attitudes toward a visionary, non-existent technology might be prone to biases and misreporting [31]. For cryptocurrency owners, these will be alleviated.

Second, tech-savvy persons are likely to be among the first adopters of the new technology.[3] This is supported by studies showing that there exists a segment of consumers who tend to adopt innovative technologies early on [2, 15, 35]. These consumers take the role of opinion leaders and influence others' attitudes or adoption decisions regarding technological products. They are characterized by a strong intrinsic affinity to high-tech, cutting-edge products and services, and are often deemed to play a special role in the process of the diffusion of innovations [34]. Hence, these persons' attitudes are informative, e. g., to assess potential initial demand for CBDC. While it is evident that cryptocurrency ownership and tech-affinity correlate, it turns out that the correlation is not as strong as one might think—most tech-affine consumers do not own cryptocurrencies. This allows us to separately analyze both tech-savviness and cryptocurrency ownership.

Third, cash still accounts for a large share of payment transactions in many advanced economies [18]. The payments literature has established that cash use is largely driven by consumers' preferences: cash is used for its low costs, for convenience, for its simplicity, for expenditure control, and to preserve privacy [36]. There are two main competing conjectures about how cash-affine consumers may view CBDC. On one hand, it is well conceivable that cash-affine people will not have a demand for a (new) digital payment instrument—simply because cash fulfills their needs. On the other hand, CBDC may as well be attractive to cash-affine users, in particular if it is convenient, generates low costs, and resolves the concerns that might have stopped them from adopting digital payments offered by the private sector [24].

The three consumer types are measured with the dummy variables *Cash-affine*, *Tech-savvy* and *Cryptocurrency owner*, respectively. Appendix B presents a definition of all variables and Table C.1 reports descriptive statistics. In our sample, 8% are cryptocurrency owners, 15% are tech-savvy and 35% are cash-affine. While the groups are intentionally not disjoint, as visualized in Figure 1, the correlation between these three groups is rather low such that we can include all three dummies simultaneously.

Further controls In order to account for confounding effects, we consider a number of basic socio-economic controls. Moreover, we include a set of background variables that could potentially have implications on respondents' attitudes towards CBDC: the stated importance of retaining cash for anonymous payments, the stated importance of hoarding cash, and trust in the central bank. To rule out that the latter variable merely reflects whether a person

[2] The questionnaire in German is available from the authors upon request.

[3] The take-up of financial innovations or digital services by tech-savvy persons is substantially higher than that of non tech-savvy persons. As a case in point, unpublished survey data shows they are about three times more likely to use alternative payment services providers like Apple Pay or Google Pay or mobile apps to send/receive money to/from persons.

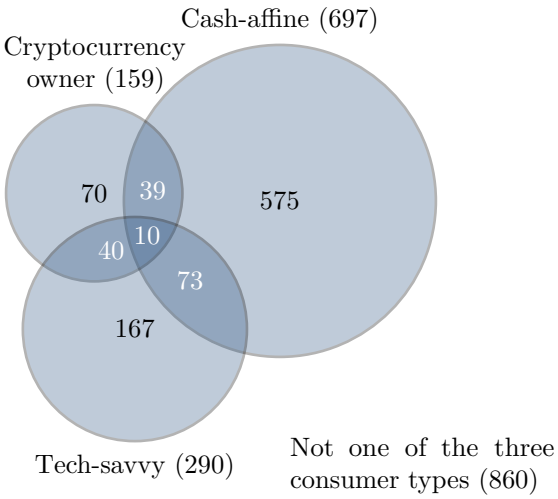


Figure 1: Venn diagram for the three consumer types (in absolute numbers). The total number of observations for each type is provided in parentheses.

is generally less or more trusting, we also include a variable measuring trust in people.

3.3 Specification

For each binary dependent variable of interest Y_i, we estimate a series of multivariate logistic regression models, specified in the basic form as

$$P(Y_i = 1|X_i) = \frac{\exp(X_i\beta)}{1+\exp(X_i\beta)}, \qquad (1)$$

where X_i denotes the row vector of respective control variables. We dichotomize individual responses reported on ordinal scales to a binary outcome following predefined rules. For compactness, each table in Appendix reports results from four regression specifications run separately for each of the two dependent variables. In the default specification (specification 1, respectively 5), X_i consists of a constant term and the three consumer types defined above. This default specification is extended with binary control variables in three steps. Specification (2, resp. 6) adds a set of socio-economic controls. This specification is fitted *without* the consumer types. Specification (3, resp. 7) combines the consumer types and the socio-economic controls. Specification (4, resp. 8) additionally includes the behavioral controls of interest (*hoarding of cash important*, *anonymity of cash important*, *trust in central bank*, *trust in people*). Occasionally, special controls are included for selected dependent variables and discussed in the respective sections below.

The logistic regression models are fitted with the maximum likelihood method. For the sake of interpretability, we refrain from reporting raw logistic regression coefficients. Instead, we calculate the average marginal effects and test their statistical significance. The coefficient values indicate the average percentage points change in the dependent variable if the binary predictor changes from zero to one. Each table also reports means of the dependent variable (which may vary across specifications due to a list-wise exclusion of missing values), the number of cases, and two goodness-of-fit measures.

Empirical studies of cryptocurrency users [1, 9, 37] find an interest in the technology to be one of the prime reasons for cryptocurrency ownership. This would suggest that cryptocurrency owners are rather similar to tech-savvy consumers. A smaller fraction of cryptocurrency ownership, however, has been found to be driven by other considerations, like the independence from banks, the idea of decentralized finance, etc. This would suggest that cryptocurrency owners have different attitudes towards money than tech-savvy persons. To test whether the respective coefficients of *Cryptocurrency owners* and *Tech-savvy* differ statistically, we report results from a likelihood ratio test (LRT) for the null hypothesis that the two coefficients are equal. The test statistic is computed from the underlying logistic model and we report the p-value for each specification where it applies.

Each regression analysis deliberately uses similar specifications. This approach inhibits the search for statistically significant effects and limits potential model selection bias.

4 Results

Before presenting the results, we note that the survey module on the digital euro was introduced by a general and simplified explanation of the digital euro. It was explicitly stated that a digital euro would be complementary to cash and that one digital euro would have the same value as one euro in cash. Respondents were told that digital euro payments would be free of charge, secure, and convenient.

4.1 Interest in CBDC

We first assess people's principal interest in the digital euro. Overall, we find that 17% of the sample

express an explicit interest and 37% state that their interest is rather limited (in the subsequent regressions these two categories are collated). 46% of the sample is not interested at all.

Column 2 in Table C.4 shows that interest is significantly higher among younger, higher income, and higher educated respondents. These findings largely mirror the results of other studies on the adoption of financial technologies [36, 8]. We find strong differences between our consumer types: tech-savvy respondents are 23 percentage points (pp) more likely to be interested and cryptocurrency owners are 15 pp more likely to be interested (column 1) than the respective comparison groups, confirming our presumption that these two groups are open-minded to the new technology. In contrast, cash-affine consumers are 29 pp less likely to be interested than non cash-affine ones. In column 3, we include socio-demographic controls and our type variables jointly. The respective results are qualitatively similar, which shows that the differences across type variables are not driven by socio-demographic factors.

To control for further confounding effects, specification 4 includes a set of additional variables: the stated importance of hoarding cash and making anonymous cash payments, as well as trust in the central bank. These variables enter significantly with the expected signs. The point estimate for *Cash-affine* is reduced slightly. Qualitatively, however, the finding that cash-affine users have a much lower interest in CBDC remains unchanged. This corroborates results from the payment literature which shows that cash users tend to react to payment innovations only sluggishly.[4] Our results indicate that this reaction is unaffected by whether the innovation is a new payment card, for example, issued by a private entity, or a new form of money issued by a central bank.

The results in Table C.4 (i. e., specifications 1–4) are of significant importance for the remainder of this paper as most of the **subsequent analyses are based on the subsample of persons reporting at least some interest in the digital euro**. This avoids noise in the data which would arise if persons who are completely uninterested in the digital euro were asked for their attitudes and preferences.

Our focus on this smaller sample introduces some changes in its key characteristics. About two thirds of cash-affine users are not interested in a digital euro. In contrast, more than 70% of tech-savvy persons and cryptocurrency owners are interested. Table C.2 contrasts the sample characteristics for interested (column 2) and uninterested persons (column 1). For almost all variables we find significant and often sizable differences, e. g., the sample we analyze henceforth is characterized by a substantial underrepresentation of cash-affine users, older persons, persons who prefer anonymous payments, persons for whom hoarding of cash is important, and risk averse persons. In contrast, there is a strong overrepresentation of tech-savvy persons, cryptocurrency owners, young, higher educated, high income persons, and of those who trust in central banks and people.

[4] For example, [14] show that payment behavior of (intensive) cash users is barely affected by the availability of contactless debit cards.

4.2 The future of cash

A common thread in policy discussions on CBDC is the question whether a CBDC may complement or substitute cash. Cash payments have declined in many countries over the past couple of years, with Sweden or Norway being known as forerunners in the transition to a cashless society [20]. CBDC could potentially accelerate this shift. We asked all survey respondents whether they believe that cash should keep its current relevance or whether it can lose importance or disappear altogether. Overall, 64% of the respondents state that cash should retain its current relevance.

Table C.4 reports the logistic regression results (specifications 5–8). Consistent with the literature [14, 23, 36], older consumers value traditional experiences and, as a result of their technology inertia, strongly advocate for the retention of cash payments. Persons with higher education or income tend to accept a decline in the relevance of cash (column 6). The effect fades out as the consumer types and other behavioral controls are added (columns 7 and 8). Unsurprisingly, cash-affine users are much more likely, whereas tech-savvy persons and cryptocurrency owners are much less likely to state that "cash should keep its current relevance." These results show that cash-affine users not only tend to oppose a digital euro, but also want cash to remain important. Some drivers for this, included in the specification 8, turn out to have strong effects. People who state that cash is needed to make anonymous payments are 26 pp more likely to support the relevance of cash. The importance of hoarding cash adds 21 pp. On average, people who agree to both reasons support the retention of cash almost unanimously. While tech-savvy respondents have a significantly lower support for

cash than the comparison group, on average, the share supporting cash is still above 50%. The same holds for cryptocurrency owners. These results empirically underpin the approach of central banks to offer CBDC as an additional offer to consumers such that cash will not be replaced.[5]

4.3 Account or token-based access

Considering the implications and path dependencies emerging from the choice of an access model, it is of interest to find out which option is more preferred by the general public. Two idealized access models, account and token-based, were presented to respondents in simplified scenarios – using the analogy of debit card and cash payments and avoiding any technical jargon. Since it is not trivial to present these choices to respondents and question wording may affect responses, Figure 2 displays the formulation of questions and the respective answers. Specifically, we have used a sequence of three questions to introduce the trade-off to the respondents. The answers show that an account-based digital euro is preferred to a token-based system (50% versus 23%). 15% of respondents have no clear preference and 13% answer that they don't know. The support for an account-based implementation is also found in the sub-populations of cash-affine users, tech-savvy persons, and cryptocurrency owners.

For the logistic regressions we have constructed a dummy variable which is 1 if respondents are in favor of a cash-like (token-based) system and 0 if they are in favor of an account-like system or if they do not care.[6] The results presented in Table C.5 show that the token-based access model is significantly less likely to be endorsed by female and older repondents. Cash-affine persons are more likely to prefer digital tokens (column 3), however the difference of 7 pp is not large enough to make the majority of this group to support a cash-like CBDC. Cryptocurrency owners show the strongest support (in relative terms) for a token-based model, which we explain with their greater familiarity with this access mode.

Typically, females and older persons are found to be more risk averse than the average consumer [22]. Specification 4 includes a dummy variable *Risk averse* which is 1 if a person is not willing to accept any financial risks in exchange for a higher than average return (see the Appendix for a definition of variables), which applies to 50% of the population. The results show that risk averse persons have a lower preference for a token-based system, on average. Controlling for risk aversion also moderates the effect of gender, age and cash affinity, as expected. Finally, column 4 includes trust in the central bank. The results show that the preference for an account model increases with the amount of trust in central banks.

Taken together, our results indicate that an overwhelming share of the population has a preference for an account-like CBDC. This applies to cash-affine consumers, tech-savvy persons, and cryptocurrency owners. This finding is connected to the risk of financial losses with risk averse persons being significantly more likely to prefer an account-based access model.[7]

4.4 Offline and P2P payments

We also asked the respondents about their preferences on selected features that have been brought up in policy discussions on CBDC design. One example is the perceived importance of making offline payments. About 40% of the respondents stated that offline functionality is "very important," another 33% considered it "important," 11% "rather not important," and only 8% "not important at all." 7% responded that they do not know. We construct a dummy variable taking a value of 1 for the first two categories and 0 otherwise (omitting don't knows).

Table C.6 (columns 1–4) reports the logistics regression results. Tech-savvy users are 12 pp more likely to consider offline functionality important than the reference group. Given the high mean of the dependent variable across the sample, this suggests that this consumer type overwhelmingly regards an offline option as indispensable. In other words, even tech-savvy persons do not believe that the internet connectivity can always be taken for granted in all future payment situations. By contrast, cash-affine users are at least 6 pp less likely to consider offline features of CBDC important. While this decline is modest against the high mean value, the result might reflect that cash-affine users have already an offline payment instrument in use.

[5] E. g., "The digital euro would not replace cash", ECB President Lagarde (https://www.ecb.europa.eu/press/key/date/2022/html/ecb.sp220114~fe1e70ec1a.en.html).

[6] Omitting don't know answers does not affect the regression results qualitatively.

[7] Although we are confident about these general findings, we note that answers are likely biased in the direction of an account-based CBDC as the questions emphasize the risk of financial losses. In future implementations of such surveys, it would be interesting to implement survey experiments with different formulations.

Q Cash-like digital euro

"Suppose that the digital euro works very similar to cash. Payments are not linked to your identity and are hard to trace. However, in case you lose such a digital euro or if you fall victim to theft, the monetary loss is irrevocable. Under such conditions, would you use a digital euro?"

Q Account-like digital euro

"And now suppose that the digital euro functions like a debit card with an account. Such payments can be linked to your identity and are traceable, but the risk of loss is very low. Under such conditions, would you use a digital euro?"

%	Would use cash-like	Would use account-like
Yes, certainly	10	15
Rather yes	31	45
Rather not	25	21
No, certainly not	24	8
I don't know.	10	11
	100	100

Q Preferences cash-like vs. account-like

"And which of these variants would you prefer: Would you rather disclose your identity and open an account to keep the risk of loss low, or would you prefer a cash-like digital euro?"

Identified account and thus no risk of loss	50
No account, but risk of loss	23
I don't care.	15
I don't know.	13
	100

Note: Subset of respondents who are generally interested in the digital euro.

Figure 2: Sequence of questions to elicit consumer preferences on token vs account-based access.

A similar picture emerges for person-to-person (P2P) payments. About 20% of the respondents consider the P2P functionality as "very important", 33% "important," 23% "rather not important," and 16% "not important at all;" 7% do not know. Table C.6 shows the regression results for a dummy variable that is constructed in the same way as before. Although fewer respondents, on average, consider the P2P functionality important than the offline fallback, we observe the same direction of effects for the controls. Tech-savvy persons are more likely to demand P2P functionality. Cash-affine users demand it less, confirming our interpretation above that these users cannot be "bought in" with features. Moreover, respondents who value the anonymity of cash payments are 12 pp less likely to demand P2P functionality than the reference group. Perhaps, they believe that cash is and will remain unchallenged for P2P payments, which indeed involve some anonymity in many social contexts (e. g., donations, but also bribes, which were not prompted). Most interestingly, cryptocurrency owners are 21 pp more supportive of the P2P payment feature. One possible explanation is that cryptocurrency owners in principle like the P2P functionality offered by cryptocurrencies, but it is not very useful for them in daily life as too few counterparties exist to transact with. Currently, making cryptocurrency payments is a niche application.[8] Cryptocurrency owners might expect that a CBDC would lead to a wider adoption of digital P2P payments in the general economy. As a result, they could benefit from the emerging network externalities.

In summary, our respondents consider an offline fallback relatively more important than P2P functionality. Both features will benefit tech-savvy users, in particular, but are unlikely to convince cash-affine persons to revisit their aversion against a digital euro.

4.5 Attitudes to security and privacy

CBDC design decisions regarding security and privacy are considered among the most critical for a broad acceptance of CBDC among consumers.

To inform CBDC designers, the survey elicits respondents' assessment of several basic attributes of a digital euro with the question "How important are the following attributes of a digital euro to you?" Answers were given on a scale from 1 (very important) to 5 (not important at all). Before discussing results, it should be noted that such an exercise, evidently, represents only a first attempt to eliciting user needs. As outlined above, the involved trade-offs are complex (e. g., between privacy and retention of transaction data) and it is difficult to make respondents aware of them by means of short survey questions. In addition, answers will depend on the chosen question wording. This cautions against stretching

[8] A striking 67% of cryptocurrency owners in our sample state that they have "never" used Bitcoin or other cryptocurrencies to pay for goods or services. Only 5% state that they do so "one or more times per month."

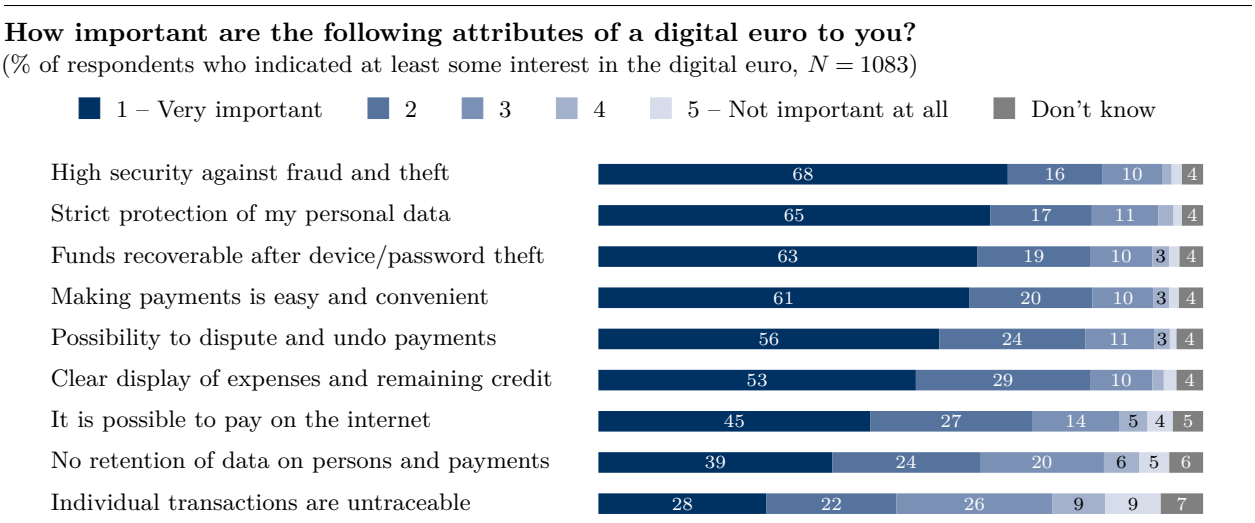


Figure 3: Importance of attributes including security and privacy items.

the interpretation of the results. Nevertheless, we consider the responses informative on how potential CBDC adopters rank security and privacy aspects.

Figure 3 summarizes the answers on all items, ranked by their importance. We have asked respondents to rate three aspects of security ("high security against fraud and theft", "funds recoverable after device/password theft", "possibility to dispute and undo payments") and three aspects of privacy ("strict protection of my personal data", "no retention of data on persons and payments", "individual transactions are untraceable"). On average, security aspects tend to be considered more important than privacy aspects. For example, almost 70% of the respondents state that high security against fraud and theft is very important. Interestingly, two privacy aspects are ranked lowest by respondents, on average. This likely reflects that most consumers have experience with electronic forms of money and have not encountered problems with the processing of respective data (e. g., by banks).

Table C.5 (specifications 5-8) reports the regression results for the leading security attribute ("high security against fraud and theft").[9] The findings show that there are no qualitative differences between cryptocurrency owners, tech-savvy persons and cash users regarding the importance of security.

Some differences are found for female and older respondents, who value security significantly more than males and the young generation (up to 35 years). In addition, risk averse persons and persons with high trust in the central bank attach more importance to security. As with regards to the consumer types, we find small significant effects, e. g., cash-affine users are 5 pp less likely to favor "high security against fraud and theft" in comparison to the reference group (specification 5 of Table C.5). Overall, these differences appear negligible in comparison to the average support for a high security against fraud and theft.

Table C.7 with results for the leading privacy attribute ("strict protection of personal data") looks very similar in terms of socio-demographic controls. Interestingly, we do not find any significant differences between the consumer types. Quite expectedly, respondents who value the anonymity of cash are more likely to emphasize the importance of personal data protection (specification 4 of Table C.7).

To rule out that the absence of significant correlations is caused by the generally high approval of these attributes, Table C.7 also presents the results for the lowest-ranked item ("individual transactions are untraceable"). We do not find any significant differences across socio-demographic variables. Among the consumer types, cryptocurrency owners are the only ones who express a sizably higher preference for untraceable transactions. This is interesting given the well-known privacy limitation of Bitcoin and other similar cryptocurrencies, in which all the transactions are public and traceable [12]. It is also

[9] We have recoded answers to a dummy variable which is 1 for "very important" and "rather important" and 0 otherwise (omitting don't know answers).

notable that cash-affine users do not endorse this privacy feature significantly more, although attitudes on the importance of the anonymity of cash as well as cash hoarding are strongly and positively associated. Perhaps only a fraction of the cash-affine respondents could link anonymity to untraceability, two non-trivial concepts, whereas many others stick to cash for convenience, conservatism, or out of habit.

5 Discussion

Our results highlight a fairly sharp divide in the perceptions of individuals who use cash more intensively as opposed to the rest. The significantly lower interest in the digital euro by cash-affine respondents is not primarily driven by socio-economic characteristics or by background variables that affect both cash-use and interest in CBDC. Overall, only slightly more than half of the respondents show some interest in the digital euro at all. Please keep in mind that the following discussion refers to this rather peculiar sub-population that has at least some interest.

User needs In terms of user needs, roughly one third of the respondents expect some advantage for themselves, should a digital euro be available as a payment instrument. Not very surprisingly and in line with [33] and [11], the responses show that younger people see more advantages than older ones. Cash-affine respondents are less likely to see an advantage from the digital euro than tech-savvy respondents or cryptocurrency owners. The sharp divide between cash-affine users and the others is corroborated by the fact that this user group not only is hesitant about the adoption of a digital euro, but also wants to see an important role for cash in the future. Note that even among the tech-savvy respondents, who are most likely to adopt a digital euro, the support for cash is strong. More than 50% wish that cash retains an important role in the future. We acknowledge that this finding may well reflect an Austrian specificity, given its still high cash intensity. Contrasting this with results from a highly cash-less society, for example, the Netherlands, could be instructive for future research.

Technology preference CBDC developers face a number of challenges when choosing specific technical implementations. The decisions often involve trade-offs. The ECB [17] presents some of these challenges in terms of principles or desiderata to be fulfilled simultaneously. Our data allow us to inform this discussion with a perspective from potential users. When confronted with simplified versions of a key trade-off, users strongly prefer an access model that resembles a bank account rather than a digital token (i.e., a digital equivalent of a bearer instrument that can get lost). This holds across all consumer types. The data suggest that risk aversion might be a key driver of this preference, which could have been amplified by the chosen question wording emphasizing the risk of losses. Despite the strength of this result, it should be considered tentative and the sensitivity to framing effects should be evaluated before deriving design decisions.

The question whether a digital euro should be equipped with offline functionality is debated among policy makers. While sometimes justified with better resilience, an offline functionality could also give a CBDC a comparative advantage over most existing forms of electronic payments. The main lesson we can learn from this study is that such a feature is regarded important by the user group we describe as tech-savvy. Cash-affine users, however, express less need for this feature, indicating that the missing offline functionality is not the main reason why they prefer cash over other available electronic payment options. When it comes to the opportunity to use the digital euro for direct payments between persons (P2P), it is overall regarded as less important compared to the offline functionality. Among the three consumer types studied, cryptocurrency owners are most supportive for a P2P functionality of a digital euro. In summary, although offline and P2P functions could make the digital euro more cash-like, offering these features would not be sufficient to convince cash-affine users to adopt it.

Security and privacy preference The public consultation by the ECB [19] revealed that security and privacy of transactions are high priority issues (for the participants of the consultation). Of course, such a consultation is prone to selection bias, which is not easy to correct for. We asked potential users about specific security and privacy concerns. Overall, our respondents seem to attribute more importance to security than to transaction privacy. This is in some contrast to the findings of the ECB consultation but concords with the results in [27]. Note that we stressed in our questionnaire that physical cash will remain available. Arguably, our respondents see no need for CBDC to provide privacy in payments.

Policy implications The debate and the feeling of urgency for the provision of CBDC in policy circles is not mirrored in the broader population: many people have not heard about a digital euro at all, and many respondents show no interest. In the group that shows some interest, we see a marked division between cash-affine users and the rest. Cash-affine users seem difficult to buy in to the idea of a retail CBDC. So, the most likely early adopters are among tech-savvy users and cryptocurrency owners. These two groups, however, do not always share the same views with respect to some key design considerations. In terms of implementation, users seem to prefer an account-like solution and be surprisingly (to the authors) indifferent with respect to transaction data privacy. These conclusions need, however, qualifications, which we have provided in the text.

CBDC designers should at least be aware how new, unknown, and exotic their considerations are to the general public. To minimize the risk of retail CBDC becoming an unsuccessful government project, they are advised to extensively and clearly explain the design options and trade-offs to the group of prospect adopters. Perhaps, the most general lesson emerging from this study is that CBDC designers cannot hope to get very precise guidance on the key design decisions by just asking users about a payment instrument that does not yet exist and that necessarily appear a bit elusive and mysterious. The development of a retail CBDC will need an intensive interaction and dialogue with prospective users. This involves monitoring the effectiveness of communication activities as well as collecting information about prevailing concerns with repeated empirical studies and methods that are robust to selection bias.

Limitations The pandemic situation forced us to use a new sampling procedure relative to prior OeNB Barometer surveys. As a result, we have limited information on the non-response bias. Both the initial contact via telephone and a rather high share of self-selected CAWI interviews cause uncertainty. To compensate for this, we checked for potential biases in relevant variables by benchmarking against external data sources and past OeNB Barometer surveys (see Appendix A). Our sample seems somewhat biased with respect to internet use, financial market participation, and risk appetite in financial investments. These variables are likely to be correlated with the willingness to adopt CBDC. We took two measures to account for this uncontrollable bias. First, we only present unweighted results.[10] Second, when discussing aggregate results, we refer to the "sample" and not to the "population." The potential sample bias is less problematic when discussing results from our multivariate analyses because we control for variables that are correlated with internet use, financial market participation, and risk attitudes.

Concerning validity, our survey demanded a lot of imagination from its participants as we were interviewing about a hypothetical technology many of them knew nothing about. Moreover, the data quality hinges on the instrument design, specifically the wording of questions. For many concepts we could not draw on established constructs and scales. While we tried to evade all avoidable pitfalls with extensive pretests, and are generally confident in the results given their coherence and plausibility, some potential framing effects cannot be fully ruled out. Finally, Austria is a small country with comparatively high cash use. Not all results from Austrian consumers might generalize to more cashless societies or the euro area as a whole.

6 Conclusion

Our empirical results suggest that it is far from certain that the introduction of a digital euro will unconditionally lead to its widespread adoption. While we provide some concrete guidance for CBDC design, we interpret our findings as tentative and exploratory.

While the scope of the data collection should be extended beyond Austria, either to the euro area as a whole or to a selected set of countries with distinct payment conventions, it is important to keep in mind that some of the assumed *social* benefits of CBDC, such as stability and privacy, seem incredibly hard to evaluate with direct questions to potential users. Innovative (combinations of) empirical methods are needed to collect valid and generalizable evidence that speaks to these questions. Our approach to identify consumer types that are more experienced with specific aspects than the general population may be worth retaining in such studies. Scaling up this research requires a lot of effort, which is worthwhile given the strategic importance attributed to retail CBDC and the strong path dependencies inherent to its technical and economic design.

[10] Post-stratification weights are available. Qualitatively, the use of weights has only a minor impact on reported percentages. See Table C.1 in the appendix for a comparison of weighted and unweighted sample means.

Acknowledgements

This research would not be possible without the time and effort devoted by thousands of anonymous survey respondents. The authors also thank Beat Weber, the discussant and participants of the Economics of Payments XI conference, and the anonymous referees for very helpful comments. Verena Fritz and Leonid Risteski provided excellent research assistance. Authors Abramova and Böhme thank the Anniversary Fund of the Oesterreichische Nationalbank for supporting their research project on "Privacy and the Functions of Digital Money."

References

[1] Svetlana Abramova, Artemij Voskobojnikov, Konstantin Beznosov, and Rainer Böhme. Bits under the mattress: Understanding different risk perceptions and security behaviors of crypto-asset users. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery.

[2] Irma Agárdi and Mónika Anetta Alt. Do digital natives use mobile payment differently than digital immigrants? A comparative study between generation X and Z. *Electronic Commerce Research*, forthcoming, 2022.

[3] Itai Agur, Anil Ari, and Giovanni Dell'Ariccia. Designing central bank digital currencies. *Journal of Monetary Economics*, 125:62–79, 2022.

[4] David Andolfatto. Assessing the impact of central bank digital currency on private banks. *The Economic Journal*, 131(634):525–540, 2021.

[5] Raphael Auer and Rainer Böhme. The technology of retail central bank digital currency. *BIS Quartely Review*, pages 85–100, March 2020.

[6] Raphael Auer, Rainer Böhme, Jeremy Clark, and Didem Demirag. Mapping the privacy landscape for central bank digital currencies. *Communications of the ACM*, 66(3):46–53, 2023.

[7] Raphael Auer, Giulio Cornelli, and Jon Frost. Rise of the central bank digital currencies: drivers, approaches and technologies. Working Paper 880, BIS, August 2021.

[8] John Bagnall, David Bounie, Kim P. Huynh, Anneke Kossed, Tobias Schmidt, Scott Schuh, and Helmut Stix. Consumer cash usage: A cross-country comparison with payment diary survey data. *International Journal of Central Banking*, 12(4):1–61, 2016.

[9] Daniela Balutel, Marie-Helene Felt, Gradon Nicholls, and Marcel Voia. Bitcoin awareness, ownership and use: 2016–20. Staff Discussion Paper 2022-10, Bank of Canada, April 2022.

[10] Bank of Canada, European Central Bank, Bank of Japan, Sveriges Riksbank, Swiss National Bank, Bank of England, Board of Governors, and Bank for International Settlements. Central bank digital currencies: foundational principles and core features. Task force report, Bank of Canada and European Central Bank and Bank of Japan and Sveriges Riksbank and Swiss National Bank and Bank of England and Board of Governors and Bank for International Settlements, October 2020.

[11] Michiel Bijlsma, Carin van der Cruijsen, Nicole Jonker, and Jelmer Reijerink. What triggers consumer adoption of CBDC? Working Paper 709, De Nederlandsche Bank, April 2021.

[12] Rainer Böhme, Nicolas Christin, Benjamin Edelman, and Tyler Moore. Bitcoin: Economics, technology, and governance. *Journal of Economic Perspectives*, 29(2):213–238, 2015.

[13] Michael D. Bordo and Andrew T. Levin. Central bank digital currency and the future of monetary policy. Working Paper 23711, National Bureau of Economic Research, August 2017.

[14] Martin Brown, Nicole Hentschel, Hannes Mettler, and Helmut Stix. The convenience of electronic payments and consumer cash demand. *Journal of Monetary Economics*, 130:86–102, 2022.

[15] Gordon C Bruner and Anand Kumar. Gadget lovers. *Journal of the Academy of Marketing Science*, 35(3):329–339, 2007.

[16] Markus Brunnermeier and Jean-Pierre Landau. The digital euro: Policy implications and perspectives. Policy report, European Parlament, October 2022.

[17] ECB. Report on a digital euro. Task force report, European Central Bank, October 2020.

[18] ECB. Study on the payment attitudes of consumers in the euro area (SPACE). https://www.ecb.europa.eu/stats/ecb_surveys/space/html/index.en.html, 2020.

[19] ECB. Eurosystem report on the public consultation on a digital euro. Consultation report, European Central Bank, April 2021.

[20] Walter Engert, Ben Fung, and Björn Segendorf. A tale of two countries: Cash demand in Canada and Sweden. Staff Discussion Paper 2019-7, Bank of Canada, August 2019.

[21] Shayan Eskandari, Jeremy Clark, David Barrera, and Elizabeth Stobert. A first look at the usability of Bitcoin key management. In *Proceedings of the 2015 Workshop on Usable Security*, 2015.

[22] Armin Falk, Anke Becker, Thomas Dohmen, Benjamin Enke, David Huffman, and Uwe Sunde. Global evidence on economic preferences. *The Quarterly Journal of Economics*, 133(4):1645–1692, 2018.

[23] Michael Harris, K Chris Cox, Carolyn Findley Musgrove, and Kathryn W Ernstberger. Consumer preferences for banking technologies by age groups. *International Journal of Bank Marketing*, 34(4):587–602, 2016.

[24] Kim P. Huynh, Jozsef Molnar, Oleksandr Shcherbakov, and Qinghui Yu. Demand for payment services and consumer welfare: The introduction of a central bank digital currency. Staff Working Paper 2020-7, Bank of Canada, March 2020.

[25] Charles Kahn, Francisco Rivadeneyra, and Tsz-Nga Wong. Eggs in one basket: Security and convenience of digital currencies. Staff Working Paper 2021-6, Bank of Canada, January 2021.

[26] Charles M. Kahn, Maarten van Oordt, and Yu Zhu. Best before: Expiring cbdc and loss recovery. Staff Working Paper 2021-67, Bank of Canada, 2021.

[27] Kantar Public. Study on new digital payment methods. Kantar Public - commissioned by the European Central Bank, 2022.

[28] Todd Keister and Daniel Sanches. Should central banks issue digital currency? *The Review of Economic Studies*, 90(1):404–431, 2022.

[29] Jiaqi Li. Predicting the demand for central bank digital currency: A structural analysis with survey data. 2021.

[30] Alexandra Mai, Katharina Pfeffer, Matthias Gusenbauer, Edgar Weippl, and Katharina Krombholz. User mental models of cryptocurrency systems—a grounded theory approach. In *Proceedings of the Sixteenth Symposium on Usable Privacy and Security (SOUPS 2018)*, pages 341–358, (virtual), 2020.

[31] Maiden Labs. Centering users in the design of digital currency. Report, MIT Digital Currency Initiative, 2021.

[32] Easwar Vivek Mangipudi, Udit Desai, Mohsen Minaei, Mainack Mondal, and Aniket Kate. Uncovering Impact of Mental Models towards Adoption of Multi-device Crypto-Wallets. Cryptology ePrint Archive, Paper 2022/075, 2022. [Accessed: 2023-06-02].

[33] OMFIF. Digital currencies: A question of trust. Report, Official Monetary and Financial Institutions Forum, 2020.

[34] Riccardo Reith, Christoph Buck, Dennis Walther, Bettina Lis, and Torsten Eymann. How privacy affects the acceptance of mobile payment solutions. In *Proceedings of the 27th European Conference on Information Systems (ECIS)*. Association for Information Systems, 2019.

[35] Riccardo Reith, Maximilian Fischer, and Bettina Lis. How to reach technological early adopters? An empirical analysis of early adopters‘ internet usage behavior in Germany. *International Journal of Innovation and Technology Management*, 17(02):2050010, 2020.

[36] Oz Shy. Cash is alive: How economists explain holding and use of cash. *Journal of Economic Literature*, forthcoming, 2022.

[37] Helmut Stix. Ownership and purchase intention of crypto-assets – survey results. *Empirica*, 48(1):65–99, 2021.

[38] Artemij Voskobojnikov, Oliver Wiese, Masoud Mehrabi Koushki, Volker Roth, and Konstantin Beznosov. The U in crypto stands for usable: An empirical study of user experience with mobile cryptocurrency wallets. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–14, New York, NY, USA, 2021. Association for Computing Machinery.

A Data quality check

Given the break in the survey mode, the fact that we have very little information on non response bias due to the contact via telephone and a rather high share of self-selected CAWI interviews, we checked for potential biases in relevant variables by comparing against external data sources and past OeNB Barometer surveys.

With respect to region, age, gender, education and income our sample is comparable with samples from previous OeNB Barometer surveys. Also employment status is comparable with a slightly lower share of retired individuals when compared to past surveys. The unweigthed sample has a slightly higher share of highly educated individuals if compared to past surveys.

As regards internet use, 94 % of individuals report that the use the internet privately on a regular base. This number can be checked against two external sources: the Austrian Internet Monitor (AIM, 2021/1)[11] with 90% and a survey by Statistics Austria from 2020 with 92%.[12]

Further splitting internet use across subpopulations, we find that internet use is 95 % among men and 93 % among women. In comparison AIM reports 94% and 87%. In our sample 80% of individuals report daily use of the internet. This compares to a rate of 80% at AIM. While the comparison suggests that our sample is by and large representative, we see nevertheless a bias in the joint consideration of age and internet use. In the age group 20-59 the gap between the OeNB-Barometer and AIM is minor with respect to internet use, the gap increases if we look at the age group above 60. For example in our sample internet use in the group 60-69 is 96% compared to 83% at AIM. For individuals above 70 we have 74% and AIM has 57%. We therefore must take into account that our sample is biased with regards to internet use in general and in particular when we look at older internet users. We suspect that the participation rate among the individuals who chose the online option is higher than for those who could give an interview by phone only.

With respect to risk attitudes we see that in the OeNB Barometer 2018 and 2019 55% reported zero risk tolerance with respect to financial decisions. In our sample the comparative rate is 50%. In past waves, 14% reported that they would accept a higher risk for a higher expected return. In our sample this share is 20%.

A direct comparison with respect to ownership of financial products is not possible since external data refer to households whereas the OeNB Barometer refers to individuals. If we take the third wave of the Household Finance and Consumption Survey (HFCS) as a reference point, we can say that 86% of households had a savings account, a life insurance or a home loan and savings contract.[13] Our sample has 79%. According to HFCS 5% of households in Austria hold stocks. In our survey 13% of individuals report stock ownership. In the OeNB-Barometer 2020/2 which was conducted with mixed methods also the stock ownership rate was 8% and thus nearer to the HFCS numbers.

Finally, we note that 8% of respondents state that they own cryptocurrencies. Previous surveys from 2019 report an ownership rate of about 2% ([37]). We consider it likely that ownership has increased from 2019 to 2021. The finding that about 40% of respondents state that they hold less than 1,000 euro in cryptocurencies suggest an inflow of new investors. Nevertheless, the ownership seems rather high when comparing with international surveys. For example, in the U.K. ownership was estimated to be 4.4% (Source: Financial Conduct Authority, 2021).

B Description of variables

Cryptocurrency owner: Derived from two survey questions. The first question asks whether respondents have heard of "Bitcoin or of other so-called cryptocurrencies". For those respon-

[11] Source: INTEGRAL Markt- und Meinungsforschungsges.m.b.H. https://www.integral.co.at/downloads/Internet/2021/07/AIM-C_1HJ21.pdf.

[12] Source: Statistik Austria https://www.statistik.at/web_de/statistiken/energie_umwelt_innovation_mobilitaet/informationsgesellschaft/ikt-einsatz_in_haushalten/index.html.

[13] See https://www.hfcs.at/ergebnisse-tabellen/hfcs-2017.html.

dents that have heard of cryptocurrencies, a follow-up question elicits the degree of interest in cryptocurrencies. Dummy variable = 1 for answers "I currently own Bitcoin" and "I currently own other cryptocurrencies", 0 otherwise.

Tech-savvy: Based on the following question: "How would you assess yourself in relation to technological developments, e.g. new devices or applications? Which of the following statement best applies to you?" Answers comprise "a) Highly interested, I would like to try new devices or applications immediately", "b) I am interested, but would not want to buy or try new devices or applications immediately", "c) I buy new devices or applications only if I see a benefit", "d) I am not interested in technological developments and only buy new devices when I need them". Tech interest high = 1 if respondents choose answer a, 0 otherwise.

Cash-affine: Derived from self-stated payment behavior. "If you think about all your purchases, including those made online, for food, clothing, services, gasoline, etc. Do you spend more (by value) in cash or more cashless – with cards or cell-phone?". Dummy variable=1 if "exclusively cash" and "more cash than cashless", 0 if "about equal", "more cashless than cash", "predominantly cashless."

Age: Measured by three dummy variables *Age group 16–35*, *Age group 36–65* and *Age group 66+*.

Female: Binary variable coded 1 for female respondents and 0 otherwise.

Urban: Dummy variable which takes a value of 1 if a respondent reports to reside in a municipality with 20,000 and more inhabitants.

High net income; income NA: Dummy variables. For those respondents who provided an answer about their household income, we compute tercils. *High net income* is coded 1 for respondents with a reported household income in the highest tercil (3.750 euro, mid-point of income brackets). *Income NA* is coded 1 for respondents who did not provide their household income (about 21% of the sample).

Academic: Dummy variable which encodes respondents with university education (1) and with non-academic background (0; e. g., mandatory schooling or technical colleges).

Hoarding of cash important: Based on "There are many people who like to have more cash at their disposal than would be necessary for daily life, as a reserve or to save. How important is it for you personally that one can hold a higher amount of cash?" Dummy variable coded as 1 for "very important" and "important", 0 for "rather not important" & "not important at all".

Anonymity of cash important: Based on the statement "Cash should be retained such that anonymuous payments can be made". Dummy variable coded as 1 if respondents "fully agree" or "somewhat agree", 0 if "somewhat disagree" and "fully disagree."

Trust in central bank: Based on "How much do you trust the following institution ... the Österreichische Nationalbank" (Central Bank of Austria)? Dummy variable coded as 1 if "very high" and "high", 0 if "rather low" or "very low".

Trust in people: Based on "Generally speaking, would you say that most people can be trusted or that you cannot be too careful in life?". Answers range from 0 ("cannot be too careful") to 10 "most people can be trusted," linearly rescaled to the unit interval.

Risk averse: Based on the question: "If there are financial decisions in your household: which of the following statements best describes your attitude toward risk: a) if I can expect a substantial profit, I am willing to take substantial financial risks; b) if I can expect an above-average profit, I am willing to take above-average risks; c) if I can expect average profits, I am willing to take average financial risks; d) I do not want to take any risk." *Risk averse* = 1 if respondents choose answer d), 0 otherwise.

C Statistical tables

Table C.1: Descriptive statistics

Variable	N	Mean value	
		unweighted	weighted
Panel A. Dependent variables (full sample)			
Interested in the introduction of a digital euro	2006	0.54	0.52
Panel B. Dependent variables (interested in the offer of a digital euro = 1)			
Belief that the digital euro brings personal advantages	914	0.43	0.43
Cash should keep its current relevance	1078	0.48	0.48
Need for a digital euro for payments on the internet	1003	0.54	0.54
Need for a digital euro for larger payments	991	0.53	0.52
Need for a digital euro for spending when traveling abroad	962	0.53	0.53
Need for a digital euro for payments for daily grocery shopping	997	0.40	0.39
Need for a digital euro for payments in hotels and restaurants	996	0.37	0.37
Need for a digital euro for sending money to persons abroad	868	0.40	0.40
Need for a digital euro for payments to persons (gifts, tips, yard sale)	986	0.33	0.34
Need for a digital euro for hoarding/saving of money	971	0.34	0.34
Preference for a cash-like digital euro	945	0.26	0.27
Would use if cash-like	970	0.46	0.47
Would use if account-like	969	0.67	0.66
Importance of offline functionality	1007	0.79	0.78
Importance of P2P functionality	1007	0.57	0.57
High security against fraud and theft	1045	0.86	0.86
Strict protection of my personal data	1045	0.85	0.83
Funds recoverable after device/password theft	1041	0.85	0.84
Making payments is easy and convenient	1040	0.85	0.84
Possibility to dispute and undo payments	1035	0.84	0.84
Clear display of expenses and remaining credit	1035	0.85	0.84
It is possible to pay on the internet	1029	0.76	0.74
No retention of data on persons and payments	1022	0.67	0.66
Individual transactions are untraceable	1007	0.53	0.53
Trust in central bank as issuer of CBDC	983	0.81	0.81
Trust in own commerical bank as issuer of CBDC	1030	0.83	0.82
Panel C. Main explanatory variables (full sample)			
Cash-affine	1984	0.35	0.38
Tech-savvy	2006	0.15	0.14
Cryptocurrency owner	2006	0.08	0.07
Panel D. Explanatory variables (interested in the offer of a digital euro = 1)			
Cash-affine	1075	0.22	0.22
Tech-savvy	1083	0.20	0.21
Cryptocurrency owner	1083	0.11	0.10
Age group 16–35	1083	0.40	0.37
Age group 36–65	1083	0.48	0.50
Age group 66+	1083	0.12	0.13
Female	1083	0.50	0.46
Academic	1083	0.21	0.17
Urban	1083	0.50	0.53
High net income	1083	0.29	0.28
Income NA	1083	0.19	0.18
Hoarding of cash important	1046	0.51	0.51
Anonymity of cash important	1060	0.82	0.82
Trust in central bank	1000	0.74	0.73
Trust in people	1071	0.42	0.43
Risk averse	1083	0.35	0.36

Table C.2: Sample comparison: Interested vs not interested in the digital euro

	Not interested Mean (1)	Interested Mean (2)	Test of equal means p-value (3)
Cash-affine	**0.51**	0.22	***
Tech-savvy	0.08	**0.20**	***
Cryptocurrency owner	0.05	**0.11**	***
Age group 36–65	**0.57**	0.48	***
Age group 66+	**0.21**	0.12	***
Female	**0.53**	0.50	
Academic	0.12	**0.21**	***
Urban	0.44	**0.50**	**
High net income	0.20	**0.29**	***
Income NA	**0.24**	0.19	**
Hoarding of cash important	**0.68**	0.51	***
Anonymity of cash important	**0.93**	0.82	***
Trust in central bank	0.63	**0.74**	***
Trust in people	0.38	**0.42**	**
Risk averse	**0.63**	0.35	***

Note: The table shows means of variables (in rows) for the sample of uninterested respondents and the sample of interested respondents. Row-wise maxima are highlighted. Significance levels for t-tests of equal means: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Table C.3: Belief that the digital euro brings personal advantages

	(1)	(2)	(3)	(4)
Cash-affine	−0.09 *		−0.09 *	−0.09 *
Tech-savvy	0.22 ***		0.20 ***	0.19 ***
Cryptocurrency owner	0.25 ***		0.23 ***	0.26 ***
Age group 36–65		−0.14 ***	−0.11 **	−0.12 ***
Age group 66+		−0.19 ***	−0.14 **	−0.13 **
Female		−0.08 *	−0.04	−0.01
Academic		0.04	0.03	0.02
Urban		−0.01	−0.01	−0.03
High net income		0.00	−0.02	−0.03
Income NA		−0.09 *	−0.09 *	−0.08
Hoarding of cash important				−0.02
Anonymity of cash important				−0.04
Trust in central bank				0.01
Trust in people				0.31 ***
Mean dependent variable	0.426	0.426	0.426	0.434
LRT tech-savvy = crypto owner	0.098		0.149	0.055
Pseudo-R^2	0.06	0.02	0.08	0.18
Log likelihood	−584	−608	−575	−509
Observations	908	914	908	829

The table shows marginal effects from logit regressions.
Subset of respondents who report at least some interest in the digital euro.
Significance levels: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Table C.4: Results of two logistic regressions (the dependent variable is shown in a multi-column header)

	Interested in a digital euro				Cash should keep its relevance			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Cash-affine	−0.29 ***		−0.26 ***	−0.22 ***	0.34 ***		0.33 ***	0.22 ***
Tech-savvy	0.23 ***		0.18 ***	0.17 ***	−0.11 ***		−0.07 **	−0.08 **
Cryptocurrency owner	0.15 ***		0.11 *	0.14 **	−0.11 **		−0.09 *	−0.09 *
Age group 36–65		−0.18 ***	−0.15 ***	−0.15 ***		0.16 ***	0.12 ***	0.12 ***
Age group 66+		−0.27 ***	−0.21 ***	−0.22 ***		0.14 ***	0.08 **	0.09 **
Female		−0.05 *	−0.03	−0.02		0.04 *	0.03	0.02
Academic		0.15 ***	0.11 ***	0.08 **		−0.11 ***	−0.07 *	−0.04
Urban		0.06 *	0.04	0.03		−0.01	0.01	0.01
High net income		0.07 **	0.04	0.03		−0.06 *	−0.03	−0.04
Income NA		−0.06 *	−0.04	−0.04		0.06 *	0.05 *	0.03
Hoarding of cash important				−0.11 ***				0.21 ***
Anonymity of cash important				−0.11 **				0.26 ***
Trust in central bank				0.08 ***				−0.02
Trust in people				0.04				−0.08 *
Mean dependent variable	0.542	0.540	0.542	0.548	0.646	0.644	0.646	0.647
LRT tech-savvy = crypto owner	0.017		0.049	0.183	0.192		0.311	0.233
Pseudo-R^2	0.11	0.05	0.14	0.27	0.11	0.03	0.13	0.33
Log likelihood	−1237	−1310	−1194	−1013	−1148	−1254	−1126	−868
Observations	1984	2006	1984	1738	1975	1991	1975	1736

The table shows marginal effects from logit regressions.
Significance levels: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Table C.5: Results of two logistic regressions (the dependent variable is shown in a multi-column header)

	Preference for a token-based access				High security is important			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Cash-affine	0.08 *		0.07 *	0.05	−0.05 *		−0.04	−0.06 **
Tech-savvy	0.07 *		0.04	0.04	0.03		0.06 *	0.06 *
Cryptocurrency owner	0.14 ***		0.10 *	0.08	−0.04		−0.01	0.01
Age group 36–65		−0.09 **	−0.08 **	−0.05		0.13 ***	0.14 ***	0.10 ***
Age group 66+		−0.20 ***	−0.18 ***	−0.15 ***		0.11 ***	0.11 ***	0.07 **
Female		−0.11 ***	−0.09 **	−0.07 *		0.10 ***	0.10 ***	0.08 ***
Academic		0.07	0.06	0.07		0.03	0.03	0.02
Urban		0.02	0.03	0.02		0.02	0.02	0.04
High net income		0.00	0.00	−0.01		0.01	0.00	0.00
Income NA		−0.02	−0.02	−0.04		−0.03	−0.03	−0.01
Hoarding of cash important				0.00				0.02
Anonymity of cash important				0.09 *				0.04
Trust in central bank				−0.07 *				0.07 **
Trust in people				0.10				−0.02
Risk averse				−0.09 **				0.08 **
Mean dependent variable	0.261	0.259	0.261	0.259	0.864	0.863	0.864	0.876
LRT tech-savvy = crypto owner	0.014		0.044	0.108	0.131		0.118	0.326
Pseudo-R^2	0.03	0.04	0.05	0.16	0.02	0.07	0.09	0.26
Log likelihood	−527	−521	−513	−455	−409	−388	−380	−309
Observations	940	945	940	853	1039	1045	1039	929

The table shows marginal effects from logit regressions.
Subset of respondents who report at least some interest in the digital euro.
Significance levels: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Table C.6: Results of two logistic regressions (the dependent variable is shown in a multi-column header)

	Importance of offline functionality				Importance of P2P functionality			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Cash-affine	−0.06 *		−0.06 *	−0.08 *	−0.16 ***		−0.16 ***	−0.18 ***
Tech-savvy	0.12 **		0.11 **	0.12 **	0.14 ***		0.10 **	0.10 *
Cryptocurrency owner	0.01		0.00	0.04	0.21 ***		0.18 **	0.22 ***
Age group 36–65		−0.05	−0.04	−0.04		−0.16 ***	−0.15 ***	−0.15 ***
Age group 66+		−0.13 **	−0.12 *	−0.10 *		−0.30 ***	−0.27 ***	−0.26 ***
Female		0.03	0.04	0.04		−0.06 *	−0.04	−0.03
Academic		0.05	0.05	0.02		0.05	0.04	0.01
Urban		−0.01	−0.01	−0.01		0.05	0.04	0.05
High net income		−0.01	−0.01	−0.03		0.03	0.01	0.00
Income NA		−0.01	−0.02	−0.04		0.01	0.00	−0.02
Hoarding of cash important				0.02				0.05
Anonymity of cash important				0.02				−0.12 **
Trust in central bank				0.08 **				0.06
Trust in people				0.04				0.07
Mean dependent variable	0.790	0.789	0.790	0.793	0.571	0.573	0.571	0.573
LRT tech-savvy = crypto owner	0.425		0.313	0.197	0.140		0.150	0.053
Pseudo-R^2	0.02	0.01	0.04	0.15	0.04	0.04	0.07	0.18
Log likelihood	−506	−511	−499	−439	−659	−663	−640	−565
Observations	1000	1007	1000	897	1001	1007	1001	905

The table shows marginal effects from logit regressions.
Subset of respondents who report at least some interest in the digital euro.
Significance levels: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Table C.7: Results of two logistic regressions (the dependent variable is shown in a multi-column header)

	Importance of protecting personal data				Importance of transaction untraceability			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Cash-affine	−0.03		−0.02	−0.05	0.06		0.06	−0.01
Tech-savvy	−0.02		0.01	0.01	0.06		0.06	0.03
Cryptocurrency owner	−0.02		0.00	0.04	0.15 **		0.17 **	0.17 **
Age group 36–65		0.10 ***	0.10 ***	0.07 **		−0.04	−0.02	−0.03
Age group 66+		0.11 ***	0.10 ***	0.07 *		0.02	0.05	0.04
Female		0.10 ***	0.10 ***	0.07 **		−0.01	0.02	0.00
Academic		0.05	0.04	0.04		−0.01	−0.02	0.00
Urban		−0.01	−0.01	−0.01		−0.01	−0.01	−0.01
High net income		−0.02	−0.02	−0.04		−0.03	−0.03	−0.05
Income NA		0.00	0.00	−0.01		0.03	0.03	−0.01
Hoarding of cash important				0.04				0.10 **
Anonymity of cash important				0.07 *				0.23 ***
Trust in central bank				0.11 ***				−0.02
Trust in people				−0.09 *				−0.03
Risk averse				0.10 ***				−0.06
Mean dependent variable	0.848	0.846	0.848	0.849	0.531	0.530	0.531	0.533
LRT tech-savvy = crypto owner	0.664		0.952	0.610	0.023		0.016	0.016
Pseudo-R^2	0.02	0.05	0.06	0.21	0.02	0.00	0.02	0.15
Log likelihood	−442	−429	−424	−353	−685	−694	−683	−592
Observations	1039	1045	1039	929	1002	1007	1002	903

The table shows marginal effects from logit regressions.
Subset of respondents who report at least some interest in the digital euro.
Significance levels: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

"Would You Give the Same Priority to the Bank and a Game? I Do Not!"
Exploring Credential Management Strategies and Obstacles during Password Manager Setup

Sabrina Klivan[C] Sandra Höltervennhoff[L] Nicolas Huaman[L] Yasemin Acar[WP]
Sascha Fahl[CL]

[C] *CISPA Helmholtz Center for Information Security*
[L] *Leibniz University Hannover*
[W] *George Washington University*
[P] *Paderborn University*

Abstract

Password managers allow users to improve password security by handling large numbers of strong and unique passwords without the burden of memorizing them. While users are encouraged to add all credentials to their password manager and update weak credentials, this task can require significant effort and thus jeopardize security benefits if not completed thoroughly. However, user strategies to add credentials, related obstacles, and their security implications are not well understood. To address this gap in security research, we performed a mixed-methods study, including expert reviews of 14 popular password managers and an online survey with 279 users of built-in and third-party password managers. We extend previous work by examining the status quo of password manager setup features and investigating password manager users' setup strategies. We confirm previous research and find that many participants utilize password managers for convenience, not as a security tool. They most commonly add credentials whenever a website is visited, and prioritize what they add. Similarly, passwords are often only updated when they are considered insecure. Additionally, we observe a severe distrust towards password managers, leading to users not adding important passwords. We conclude our work by giving recommendations for password manager developers to help users overcome the obstacles we identified.

1 Introduction

Despite investigations into new online authentication standards [11, 17, 41, 67], usernames and passwords remain the most widely used. Users need to manage an enormous amount of online credentials, which has only increased with the growth of online communication during the recent global pandemic [34, 69]. Due to the number of accounts, users face an immense cognitive burden when creating and memorizing strong and unique passwords for all of them [23, 24, 49, 52, 70, 71, 73].

A promising way to mitigate the above challenges is the use of password managers (PWMs). They allow users to maintain all their passwords and often additional information such as credit card data, addresses, or two-factor authentication secrets behind a single master password. Users therefore only need to memorize this one password, removing most of the cognitive load [59, 63]. Most PWMs furthermore provide password security checks, the generation of strong passwords [33, 46], and provide auto-save and autofill features. However, the initial PWM setup requires a lot of time and effort: Users need to choose and install a PWM as well as potential web browser extensions, gather their online accounts, add them one by one into the PWM and ideally also update weak, re-used or leaked passwords. All of this is time-consuming and requires users to have a list of all of their accounts ready if they want to set up their PWM as quickly as possible, but composing this list is often a challenging task. On the other hand, the effective security benefits of PWM are reduced if users do not add and upgrade their credentials when adopting the PWM, as passwords might remain reused or easily guessable. In this work, we extend previous work and aim to understand what strategies users actually apply during their initial PWM setup. This includes how they add new or existing passwords, if and how old passwords are updated, users' thought processes and perceptions, and finally, which obstacles they face during the setup. Based on our findings, we give recommendations to PWM developers on how to improve the process and help PWM users with password management tasks.

We initially collect helpful features for new users when first setting up a PWM by conducting an expert review, evaluating several popular PWMs. Based on this expert review and

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.
USENIX Symposium on Usable Privacy and Security (SOUPS) 2022.
August 7–9, 2022, Boston, MA, United States.

extensive piloting to collect potential management strategies, we follow up with a survey with 279 users of built-in and third-party PWMs. To the best of our knowledge, we are the first to investigate PWM setup support features and credential management strategies users apply when setting up their PWMs.

In this work, we aim to answer the following research questions:

- **RQ1**: *What setup features do password managers offer to new users, who want to add their existing credentials?*
- **RQ2**: *What are common user strategies to add new and existing credentials? Why are these strategies used?*
- **RQ3**: *How can password manager developers help users with setup, and improve the overall process?*

Overall, we make the following contributions:

Existing Setup Features. We perform expert reviews of 14 popular PWMs and present and evaluate current built-in tools and features that help users to add new and existing credentials as quickly as possible, and to identify and update potentially weaker passwords efficiently.

Strategy Identification. We provide a first exploratory investigation, in which we identify seven PWM credential management strategies end users adopt, and obstacles they face during setup.

Frequency of Strategies and Issues. We design and conduct a survey study with 279 participants and report how common respective strategies and issues are. Furthermore, we investigate the reasons that influence users' strategy decisions.

Recommendations for Developers. Based on our findings, we give recommendations on how PWM developers could improve the setup process or aid (first time) users with the setup of their PWM, to help them improve their password strength and fully benefit from PWMs.

Replication Package Availability. To increase research transparency and allow for easier replication, we provide a comprehensive collection of our research artifacts on a complimentary website, including videos from our expert review, all text material from our survey, and additional aggregated survey results[1].

2 Basic Features of Password Managers

In the following, we present the prominent PWM functionalities, to help understand which tasks users face when initially adopting a PWM, and in which ways security and usability are influenced by them.

Store Credentials: At their core, PWMs are simple databases that store sensitive information including username and password pairs, and encrypt them using a master password.

Password Generation: PWMs can generate unique and strong passwords that end users can use when updating existing, or creating and storing new passwords. These generators often come with many options, allowing users to set length, include or exclude certain characters, and gain feedback on password strength. However, these generators can struggle with services' password policies, as websites might, e. g., not permit certain symbols and force users to manually adjust the generator's settings [26, 28, 36].

Auto-save/Auto-fill: Although this often requires separate browser extensions, PWMs can detect visited websites and login forms, and automatically save entered credentials, or match the website to a known one and automatically fill the credentials in. While this streamlines the user experience and increases usability, it requires the PWM to recognize the service as well as the login form correctly, which previous work found to be a non-trivial process [26]. Some PWMs further support fully automated logins [18, 35].

Additional Data: Many PWMs can also store additional data, e. g., addresses, credit card data, or secrets required to generate Time-Based One-Time Passwords (TOTPs). Additionally, some PWMs can not only store these secrets, but to also generate and autofill valid TOTPs [59].

Synchronization: Some PWMs offer applications on different devices, therefore enabling end users to access their passwords on multiple devices such as private or work computers, smartphones and more. In these cases, the encrypted password database is usually stored in a cloud. While some PWMs such as 1Password [1] provide fully automatic cloud synchronization, other purely offline PWMs such as KeePassXC [32] only support multiple devices if end users share or synchronize the encrypted database file with themselves.

3 Related Work

In the past, adoption sentiments as well as the usability of PWMs was researched exhaustively. We present and discuss related work in two key areas: *Motivation to Use Password Managers* and *Password Manager Usability*, and illustrate how our work extends previous studies and fills an important research gap.

3.1 Motivation to Use Password Managers

In 2016, Alkadi and Renaud collected reviews of two popular PWMs from Android and Apple app stores. Based on the user sentiments, they designed a survey and report an extensive list of reasons for and against PWM use, such as ease of use, perceived usefulness, cost, perceived effort and privacy or security concerns [4]. Similarly, in 2017, Fagan et al. conducted a survey with 137 users and 111 non-users of PWMs to understand their motivations. They find users to be mainly

[1] https://publications.teamusec.de/2023-soups-pwm-adoption/

driven by convenience and usability factors, while non-users mention security concerns [21]. In 2017, Aurigemma et al. surveyed 283 undergraduates to understand why they do not use PWMs, even if they have high intentions to do so. Their study indicates that users do not adopt PWMs due to various concerns about trust, costs or actual benefits, and that even users who are interested in PWM usage are inhibited by time constraints and a lack of immediate threats [7]. A 2018 survey conducted by Maclean and Ophoff examines adoption intentions based on technology acceptance and use, and finds the expectancy of functionality, trust into the system, and that usage becomes a habit to be leading factors [42]. Ayyagari et al. performed a survey in 2019 to investigate the low adoption rates of PWMs and report that the perceived severity of password loss consequences greatly influences end users' likelihood to use PWMs [8]. In 2021 Albayram et al. conducted a series of surveys to examine the impact of motivational text and video material about the benefits of PWMs on improving the understanding and adoption rate of PWMs. They find that both increased user comprehension, but that video material resulted in a higher adoption rate [3].

While the majority of these works researched mindsets of users that do not necessarily use PWMs, our work focuses on the experiences PWM users have when adding and maintaining passwords. Furthermore, our work provides insight into the impact of issues PWM users encounter, allowing us to determine the most important issues currently blocking the adoption of PWMs.

3.2 Password Manager Usability

Below, we discuss previous research that focuses on the usability of PWMs, as this can have a high impact on how likely end users keep or abandon a PWM. In 2006, Chiasson and van Oorschot compared the usability of two proposed PWMs in a user study. They find their participants to have strong misconceptions and conclude that not only were usability issues present, but that some of them could also lead to security problems [15]. Another usability comparison was conducted in 2010, when Karole et al. asked end users to test three different PWMs. They find that users preferred portable variants over an online PWM despite reporting lower usability, most likely due to concerns against storing passwords online [30]. Lyastani et al. conducted an in-situ examination of PWM usage and usability in 2018 and found that while PWMs increase security, the degree of this is highly dependent on a combination of password creation, storage, and entry behaviors as well as user's PWM choice [40]. In 2019, Alkadi et al. developed and distributed a recommendation app that allowed users to set several preferences and suggested the best-fitting PWM. Overall, only 5% reported installing and using it. Participants stated that the effort to set the PWM up, lack of trust and external factors such as lack of storage space are main reasons against the installation [5]. In the same year, Seiler-Hwang et al. instructed users to install a PWM on their phone and collected usability feedback with a survey. They find that even popular smartphone PWMs have severe usability deficiencies [65]. Also in 2019, Chaudhary et al. conducted a systematic literature review of 32 academic PWM proposals and examine them for usability and security. Discovering that most proposals are biased towards security and lack usability, they give recommendations for usability enhancements to PWM manufacturers [13]. Pearman et al. reported a series of semi-structured interviews in 2019, examining to what extent users utilized additional features such as strong password generation. They find that users of built-in PWMs without additional features often apply weaker passwords, and that the reasons to adopt differ between convenience for built-in PWMs and security concerns for additional installed PWMs [50]. This study was replicated in 2021 by Ray et al., with older adults. They find a higher mistrust in technologies such as cloud storage, but also motivation through family recommendations or education to be vastly more effective [57]. In 2021, Simmons et al. systematized 17 different use cases for PWMs, and performed a first usability investigation of these using cognitive walkthroughs [66]. In 2022, Oesch et al. performed 32 observational interviews to study how end users use their PWMs. They find that users are often overwhelmed or distrustful of PWMs, therefore using multiple ones as backups, and avoiding features such as, e. g., strong password generation [47]. In the same year, Zibaei et al. conducted a user study to investigate the effectiveness of secure, auto-generated password suggestions through PWMs built into Firefox, Chrome, and Safari. They find that Safari's approach to already pre-fill password fields with secure passwords led to the highest password adoption rate [76].

In contrast to the described related work, we focus especially on credential management strategies users apply to transfer their existing passwords or add new ones. While previous work discussed general user sentiments and adoption reasons, we investigate the behavior after adoption, and provide more in-depth insights into the impact of typical usability issues during PWM setup. We aim to improve the setup process and thereby overall security gain from using PWMs.

4 Password Manager Expert Review

With expert reviews of PWMs and their setup features, we aimed to answer RQ1. We were interested in features that can support users with the tedious initial setup processes when adopting a PWM, i. e., features that were designed to help them add passwords and replace them with strong alternatives where necessary to increase the security benefits from using a PWM. We used the expert reviews findings to inform our survey (cf. Section 5) and design recommendations for PWM developers. Figure 1 depicts the course of our research.

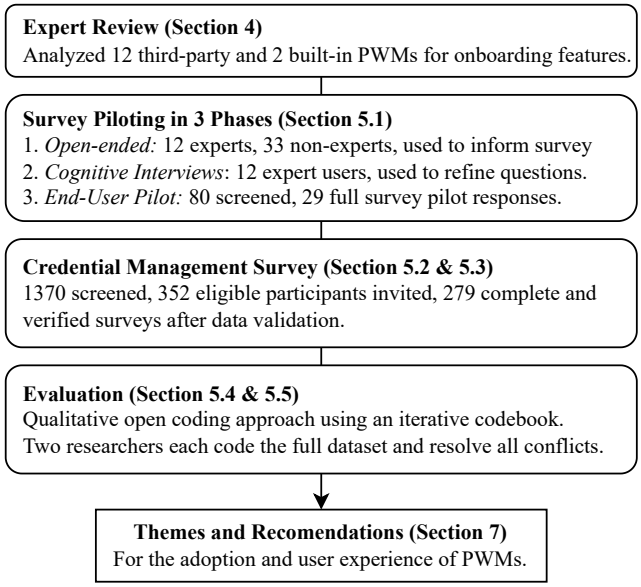


Figure 1: Overview of the methodology of our expert reviews and online survey with PWM users.

4.1 Methodology

Two authors conducted expert reviews of the 14 most popular PWMs, an approach used by previous work [20, 51] based on cognitive walkthroughs [53, 61].

We created a list of popular PWMs by collecting extension download counts for Chrome and Firefox. Since we did not aim for an exhaustive overview, and Chrome covers 65% of browser usage [68], we chose to only investigate extensions within the top ten of either. We additionally tested both browsers as well, as they offer built-in PWMs. Overall, we ended up with 14 different tools. The PWMs on this list include both offline PWMs and online PWMs with cloud storage back-ends. While we tried to use only free account plans for a better comparability, some PWMs only provided free premium trials (cf. Table 1).

For the expert reviews, we installed all PWMs on a Ubuntu 20.04. with clean Chrome profiles for each PWM. We performed a set of tasks users typically encounter after setting up a PWM, and aimed to include both common tasks, and workflows that increased security by, e. g., upgrading password strength. First, we installed the PWM including the browser extension, trying to set a bad master password to test if the PWM allowed the password that secured the remaining accounts to be weak. We followed all setup prompts or tutorials to experience every guide designed to help fresh users. Afterward, we searched for mass import features and took notes of their properties. To test in which ways the addition of account details was supported and how well users were aided in choosing strong passwords, we added several accounts. Overall, we searched for account suggestions and automated recognition of websites and their URLs, password generation features, flagging of weak, breached or reused passwords including reuse of the master password. We further searched for dedicated security centers that provide users with an overview of their account security and potentially vulnerable credentials, or for support of timed one-time passwords (TOTP) (cf. Appendix A for a full list of all tasks).

While working on these tasks, we recorded screencasts for later comparison and discussion to ensure nothing was missed and to improve the transparency of our work. Common for cognitive walkthroughs, we tried to simulate the perspective of new users by asking ourselves if the availability of features is apparent, whether the functionality and success of user actions is clearly communicated and easy to understand, and whether relevant features are present or missing. We present our findings, including the most commonly present features, in the following section.

Table 1: Overview of on-boarding features present in popular PWMs. The browser columns indicate that the respective PWM is present in the top 10 PWMs at the time of our analysis.

		Education			PW Storage				Secure PWs				
PWM	**Plan**	**Tutorial**	**Next Steps**	**Achievements**	**Bulk Import**	**Account Suggestions**	**AutoSave**	**TOTPFields**	**Security Center**	**Breach Warning**	**PW Meter**	**Enforces Secure MasterPW**	**Standalone**
LastPass[F,C]	Paid	●	●	●	●	●	●	◐	●	●	●	●	○
Bitwarden[F,C]	Free	○	○	○	●	○	◐	◐	◐	◐	○	●	●
Norton[F,C]	Free	○	○	○	●	○	●	○	●	○	●	●	○
1Password[F,C]	Paid	◐	●	○	●	○	◐	●	●	●	◐	●	●
RoboForm[F,C]	Paid	●	○	○	●	●	◐	◐	●	●	○	●	◐
KeePassXC[F]	Free	○	○	○	●	○	●	●	●	○	●	○	●
Kee[F]	Paid	●	○	○	●	○	●	●	○	○	○	○	○
NordPass[F,C]	Paid	○	●	○	●	●	●	○	◐	◐	●	●	●
Keeper[F,C]	Paid	●	○	○	●	●	●	○	●	◐	●	●	●
Avira[F,C]	Free	●	●	○	●	○	●	○	●	●	●	●	○
Dashlane[C]	Paid	●	○	○	●	○	●	○	●	◐	◐	●	○
MultiPassword[C]	Paid	○	○	○	●	○	●	●	●	○	●	◐	●
Chrome PWM	Free	○	○	○	○	○	●	○	○	●	○	◐	-
Firefox PWM	Free	○	○	○	●	○	●	○	●	●	○	◐	-
Total:		6	4	1	13	4	11	4	10	6	7	9	6

● = Feature found; ◐ = Feature conditionally found; ○ = Feature not found.
[F] = From Firefox Top 10, [C] = From Chrome Top 10

4.2 Results

Overall, our main focus was to identify relevant features for secure addition of credentials that users can benefit from during setup. We found that the majority of tools only offered free

premium trials. Where possible, we used the standalone program, which was the case for six PWMs and the two browsers. We identified three different categories of features that help users when initially adding their credentials, which we describe in the following:

User Motivation and Education. This type of feature serves as a first introduction to the PWM, and we distinguish between **tutorials** and **next steps**. A tutorial consists of simple, guided steps that are realized through, for example, pop-ups or interactive demonstrations, during which users are taught how they can work with the PWM and complete common tasks. In cases in which this process was not guided, but simply a non-enumerated list of available features and sensible next actions, we considered them as a list of next steps. We found tutorials present in half of all reviewed PWMs, while next step lists were offered in four. With LastPass, we additionally found one PWM that coupled its guides to **achievements** or badges to incentivize their completion and motivate users to get acquainted with the most important features. Users were further offered a 10% discount on premium memberships if they complete all achievements. Because this requires adding at least ten credentials, the achievement feature incentivizes active learning rather than simply reading a tutorial.

Ease of Password Storage. As the initial addition of account credentials is a time-consuming task, we found different functionalities aiming to ease the process. Most importantly, almost all tested PWMs except the Firefox built-in offered some kind of **bulk import** from other PWMs, browsers, or raw .csv files.

However, we found the expected import formats to differ widely, and we found different requirements for .csv files in terms of, e. g., required columns and data formatting. Users without a previous PWMs to import from would therefore need to create very specific files manually, while users whose previous PWMs export differs too much from the expected import format need to invest time to edit the data. As a result, this offers only a small benefit for users, who need to compile the respective file, requiring them to remember all their accounts in a similar problematic way than if they had added all credentials one by one. Additionally, this encourages them creating a non-encrypted collection of credentials they might not remove from their hard drive afterward [12,58]. A slower approach is the ability to **auto-save** passwords while browsing, in which the PWM extension offers to store credentials whenever the user logs into a website or registers new accounts. We found this available in almost all PWMs, however, it was only available as a premium feature in RoboForm. While both Bitwarden and 1Password in theory offered auto-saves, we experienced issues with this feature in both extensions. According to their forums, the Bitwarden issue has been known for a while, and they are working on a solution [10]. Finally, we found four PWMs that **suggested popular websites** when adding new accounts. This feature is useful to help users remember which accounts they might have, however, it is only occasionally offered.

Secure Passwords. Another important step of the initial PWM setup is the chance to upgrade old passwords if they are, e. g., weak or reused, which can be supported by PWMs through signalling which passwords may need to be changed. While **password meters** are the best-known features to help users create strong passwords, we only found them in seven PWMs. When included, we often perceived them as counterintuitive. In practice, changing settings such as increasing the length regenerates the password, and while longer, the new one might have a similar, but slightly decreased entropy. However, as this is often what password meters measure, the strength bar can go down when, e. g., the password length is increased. In other cases, the evaluation was performed after the entry was stored, requiring users to actively check for warnings instead of receiving them while saving the password. Two PWMs only used password meters while using the built-in generators, therefore not providing feedback to users who create manual passwords or copy and paste old ones. Other measures include **security centers**, i. e., dashboards in which users receive comprehensive summaries of insufficient credentials that are, e. g., weak, reused over multiple stored entries, generally common, or present in leaks. This enables users to purposefully upgrade insecure account credentials where necessary, and was present in a majority (ten) of evaluated PWMs. However, we found it often only available in premium account plans, and Bitwarden only offered it in its web client, with no further mention of the feature within the standalone app.

NordPass asked us to change older passwords, although research has found regular updates to have negative impacts on security [14]. A similar feature, often included within security centers, are **breach reports**, in which either the user's email address or the passwords within the PWM are scanned for their presence in credential leaks. While present in almost all PWMs, breach reports are typically a premium feature that is not accessible for non-paying customers. This is especially curious as it is often based on the free tool Have I Been Pwned [27]. In the case of Keeper, this was particularly severe, as we were informed that some of our accounts were breached, but then asked to pay to receive any further information of which account was affected.

5 Credential Management Survey

Following our expert review of PWM setup features within popular PWMs, we conducted a survey with 279 users of both built-in and third-party PWMs. We describe the methodology of our survey study below.

5.1 Survey Design & Piloting

For the initial survey design, we created an early draft of our survey and tested it with 12 usable security expert users,

and 33 non-expert users in several rounds of piloting. The exploratory survey draft consisted of an early version of our final survey. It was modified to contain only free-text questions, which we used to collect options for multiple-choice questions in the final survey and to identify credential management strategies. For the expert survey, we additionally offered text boxes on every survey page to gather expert feedback on the question design. Based on results of this early version, we modified the survey to improve question and answer phrasing, and we used the responses to open-ended questions to create options for multiple-choice versions of some questions. Whenever a participant's answer was not yet collected as a closed-ended answer option for our final survey, we added it along with any related answer that came up during the result inspection. We stopped recruitment when we reached theoretic saturation, that is, when no new answer options emerged, and no participant answered any question in a manner that indicated a lack of understanding.

To improve survey quality and explore the area further, we additionally conducted 12 cognitive interviews [54] with associates of the authors who did not yet fill the pilot survey, and were not involved in this research project. While most of them were usable security researchers, one was an end user with a master's degree in computer science. We invited participants to a voice chat and asked them to screen share their completion of our survey. We encouraged participants to "think aloud", rephrase questions in their own words or elaborate on their thoughts to learn how they understood certain questions or why they answered in a certain way. Overall, cognitive interviews are a common approach to collect feedback and improve survey quality [2,31,39,60,74]. After each interview, two authors analyzed the responses, received feedback and agreed on survey changes. The survey was adjusted before moving to the next participant, to test changes. We recruited participants for the cognitive interviews until we found no major new misconceptions or problems.

Finally, we conducted several rounds of piloting with the closed-ended version of the survey, screening a total of 80 end users, of which we invited 33 to the full survey, and received 29 answers. We polished our question phrasing, and continued until all questions were answered with sufficient quality, indicating that the survey was now easily understandable.

5.2 Survey Structure

We designed the survey to explore which credential management strategies users apply when they initially set up their PWM, i. e., how they add their passwords, and in which ways they interact with their PWM to increase task efficiency or password security. The full survey can be found in Appendix C.

We first included PWM demographics such as when they started using a PWM (Q1), what their first PWM was (Q2) and if changed, what their current PWM is (Q3), whether they paid for it (Q4) and if they would recommend it to others (Q5). We further asked whether they added all private (Q6-Q7) or work-related (Q8-Q9) accounts to it. Afterward, we asked about their reasons to use a PWM (Q10), including who recommended it, if anyone did (Q11), and if they or somebody they knew experienced a password breach (Q12), as we deemed both relevant to their decisions regarding their PWM usage.

Overall, we aimed to investigate the spread of the PWM **credential management strategies** we identified during piloting, as well as what influenced a users' decision to apply a strategy. Therefore, we asked participants about the strategy they mainly applied (Q13) both in an open-ended question to gather unbiased experiences and sentiments, and a closed-ended version on the next survey page (Q14) to better pinpoint the precise strategy. We further asked participants to describe reasons for their strategy choice (Q15) and alterations in their current strategy to account for changes over time (Q16). Since these strategies might depend on specific website (types), we gave participants the option to share these priorities with us (Q17).

Furthermore, we were interested in how participants dealt with their existing passwords - i. e., whether they changed all, some or none of them when they set up the PWM (Q18), and their reasons for doing so (Q19-Q21). We were further interested in their password generation process (Q22-23).

Additionally, we asked broader questions regarding their experiences with the setup process (Q24-25), and which additional features of their PWMs participants were using (Q26).

Based on previous answers, we determined every participant who did not use the approach we deemed ideal from a security perspective(i. e., stated to not have updated all passwords when adding them or to not have added all accounts at once), and asked them an additional question regarding their reasoning (Q27).

To further collect insights into problems and usability improvements, we directly asked participants what their PWM could have done to improve their personal setup process (Q28).

Finally, the last part of our survey covered **common demographic questions**. This includes gender (Q29), age (Q30), and ethnicity (Q31), their highest formal education (Q32) and whether the participants ever studied a computer science related subject (Q33) or held a computer science related job (Q34).

5.3 Data Collection & Recruitment

We contacted expert users for our piloting through our professional network. For our survey study, we used the crowdsourcing platform Prolific [55] to gather participants due to their general high data quality [48] and in several rounds invited 1,370 participants to our screening survey. To uphold certain quality standards, we required participants to have

a job approval rate of at least 90% and at least five previously submitted jobs, and excluded all users who participated in previous iterations of our pilot. To filter out participants who never adopted a PWM, or were unaware of the PWM functionalities within their browser or operating system, we used two questions regarding password management in our screening survey, but did not mention our focus on PWMs yet (see Appendix B). We further asked which PWM they used and since when, and used their answers to determine usage of built-in or third-party PWMs, and as an attention and sanity check between screening and full survey. From all participants we screened, we manually selected all who stated to use a third-party PWM, and a similar amount of users of built-in PWMs. Based on results from previous work, which found significant differences in the approaches and sentiments of both groups, we decided to invite equal participant numbers for both [43, 50]. This resulted in 352 participants we invited to complete the full survey, of which 304 completed it, and 279 yielded valid answers.

5.4 Data Cleaning & Analysis

We removed 25 participants whose answers contradicted their statements from the screening survey, and found none who finished the survey suspiciously quickly. In the case of open-ended questions, two researchers coded all answers using an iterative approach based on thematic analysis [16]. For each question, they individually created a codebook, in which they denoted recurring answer patterns. They discussed their individual codebooks and merged them to create a single codebook. The two researchers jointly read all answers and assigned matching codes from the corresponding codebook. In case of conflicts or changes in the codes, both researchers discussed the problem until they reached agreement. These discussions included adjustments in the definition of individual codes, and merges or splits of codes that were rarely or frequently assigned to account for nuances in answers. Both researchers revisited previously coded answers and updated the assigned codes for a consistent coding strategy. We did not calculate inter-rater reliability due to the exploratory nature of our coding, which in theory led to a perfect agreement since we solved all conflicts via discussion [44]. The final codebook with descriptions of the codes and their distribution onto the open-ended questions can be found in Appendix D. For some selected questions, we tested for significant differences between users of built-in and third-party tools using a Chi-square test (χ^2).

5.5 Results

In this section, we describe the results of our survey study with 279 participants, asking about their strategies to initially add and update passwords to their PWM. Overall, we found that most participants add credentials whenever they access the respective services. This was often motivated by efficiency, convenience, or a lack of overview over their online accounts. Due to this, security was only a secondary factor. Many participants were deterred from investing time and effort to improve their online credential security.

Table 2: Demographics for all valid participants.

Demographics	Value	Percent
Gender:		
Man	175	62.72%
Woman	103	36.92%
Genderqueer	1	0.36%
Age:		
Median	35.0	-
Mean	30.19	-
Standard Deviation	9.24	-
Ethnicity:		
White or of European descent	191	68.46%
Black or of African descent	59	21.15%
Multiple Ethnicities	15	5.38%
Hispanic or Latino/a/x	7	2.51%
East Asian	1	0.36%
South Asian	2	0.72%
Middle Eastern	3	1.08%
Southeast Asian	1	0.36%
Education:		
Bachelor Degree	108	38.71%
Master Degree	59	21.15%
Secondary School	33	11.83%
College/University Study (without Degree)	45	16.13%
Trade/ Technical/ Vocational	13	4.66%
Associate Degree	8	2.87%
Professional Degree	3	1.08%
Other Doctoral Degree	6	2.15%
Technical Background:		
Computer Science/ Technical Education	76	27.24%
Computer Science/ Technical Job	98	35.13%
Start with PWM		
In the last week	6	2.15%
In the last month	1	0.36%
In the last six months	15	5.38%
In the last two years	52	18.64%
More than two years ago	200	71.68%
I don't know	5	1.79%

5.5.1 Participant Demographics

In this section, we provide a summary of our participants' demographics (cf. Table 2). 175 (62.72%) of our participants identified as men, while 103 (36.92%) identified as women and one person self-described as genderqueer. Participants were 30.19 years old on average (std: 9.24, med: 35.0). The majority (191, 68.46%) described themselves as White or of European descent, which is not surprising for Prolific as a European crowdsourcing platform. This is followed by 59 (21.15%) who identified as Black or of African descent. Considering education, 108 (38.71%) participants stated to have a Bachelor's degree, 59 (21.15%) a Master's degree and 33 (11.83%) have completed secondary school. 27.24% of all

participants declared that they received a degree in computer science or a related field, and 35.13% stated that they have worked in this area before. We acknowledge that computing professionals are over-represented, and assume that PWM use tracks with computer science experience. Overall, our recruited sample is in line with previous studies conducted on Prolific [31, 62].

We found that the vast majority of participants (200, 71.68%) reported using a PWM for more than two years or started using one within the last two years (52, 18.64%). Overall, 141 reported to mainly use PWMs built into their operating systems or browsers, while 138 stated to mainly use third-party PWM tools. We asked participants about their first PWMs, and which one they currently used, if it had changed. We found Chrome (84, 59.57%) and Apple Keychain (43, 30.5%) to be the most common currently used PWMs for built-in users. For third-party PWM users, the distribution was more even, with Bitwarden (39, 28.26%), KeePass variants (24, 17.39%), and LastPass (22, 15.94%) being the most frequently named. Finally, 27 (9.68% of all participants) stated uncommon PWMs that were overall only named at most three times, and 3 (1.08%) had stopped using a PWM.

5.5.2 Credential Management Strategies

In this work, we were most interested in how end users insert their account credentials when setting up a PWM, as this process can be tedious, but also crucial for improved security. Overall, we identified seven main strategies. We mainly distinguish them by the time passwords were added (e. g., immediately on install, or when services are accessed) or the choice of which passwords were added (e. g., for more or less important or frequently used accounts), and provide a description of strategies as well as their frequency in Table 3. We were unable to map 12 (4.3%) participant answers to one of our identified strategies. These participants reported to, e. g., be unable to recall, not having a strategy, or not having control over the process, as it was executed by a workplace. In other cases, the answer did not allow us to concisely determine which heuristic was used to prioritize accounts that were added, and we decided to merge them in an additional *Any Priority* strategy. Overall, we are confident to have uncovered all relevant credential management strategies in our analysis.

Adding Passwords on the Fly is Easier: The most frequent initial strategy was to add accounts whenever they were accessed for the first time after the PWM setup (108, 38.71%). This often uses auto-save features, i. e., the user does not necessarily need to consciously or manually add passwords, but can simply follow a prompt. Following, users reported to have added their most important or frequently used accounts first (81, 29.03%). This was most often reported as an attempt to save time when adding accounts. Additionally, accounts that did not contain sensitive information were typically not considered worthy of securing them. Note that we merged the

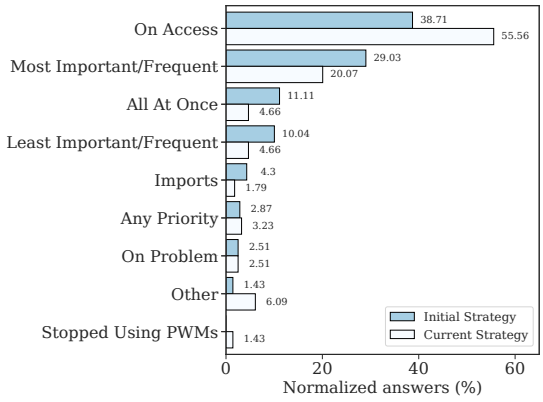


Figure 2: Distribution of the participants' initial and current main strategy in %.

strategies *most important* and *frequent* to improve reporting quality, as participants used the terms interchangeably. We found all other strategies to be less relevant in practice. From a security perspective, adding and updating all passwords at once would be ideal. However, we found only 31 (11.11%) participants to specify that their strategy was to add everything at once. We found that 28 (10.04%) users mentioned adding their least important or only rarely used accounts, often referring to security concerns or simply wanting to test the PWM before adding relevant accounts as reasons: "*I started with the less important accounts because I wanted to get used to the tool before importing my important account information.*" (P14) Other strategies our participants mentioned included importing accounts from, e. g., browsers (12, 4.3%), additions with other or unclear priorities (8, 2.87%), e. g. based on chosen passwords, or only adding accounts when a problem occurred such as forgetting the password (7, 2.51%). As these situations forced users to reset their password anyway, and to prevent the need to change it again, they decided to add it to the PWM.

In addition to their initial strategy, we asked participants to provide their current one, and to detail potential changes and reasons. We found a huge increase in users who add their credentials on access (155, 55.56%). As we were asking for the current strategy, this relates mostly to accounts that users freshly create, and that the majority of users immediately add to their PWM, presumably with the help of auto-save features. Again, the second-largest group of participants reported prioritizing important or frequently used accounts (56, 20.07%), suggesting that in these cases, lesser important accounts are never added. We found other previously mentioned strategies mostly irrelevant after the initial setup. Overall, the changes between initial and current strategy can be seen in Figure 2. Most participants chose their initial strategy due to its efficiency (113, 40.5%) or because it was perceived as the easiest approach (94, 33.69%). We found all other reasons less com-

Table 3: Participants' strategies to insert credentials into their PWM (cf. Section 5.5)

Strategy	Description	Initial	Current
All At Once	As far as practicable, users try to add all their accounts at once. Excludes *Imports*.	31 (11.11%)	13 (4.66%)
Any Priority	Users started by adding specific accounts, but described other/unclear prioritization methods.	8 (2.87%)	9 (3.23%)
Imports	Accounts were imported from, e. g., browser profiles or previous PWMs.	12 (4.3%)	5 (1.79%)
Least Important/Frequent	Users started by adding their least important and/or frequently used accounts first.	28 (10.04%)	13 (4.66%)
Most Important/Frequent	Users started by adding their most important and/or frequently used accounts first.	81 (29.03%)	56 (20.07%)
On Access	The accounts are added on the fly, whenever the account is accessed.	108 (38.71%)	155 (55.56%)
On Problem	Users enter accounts when problems occur, e. g., when they need to reset their password.	7 (2.51%)	7 (2.51%)
Other	Other strategies, mixes strategies, or unclear answers.	4 (1.43%)	17 (6.09%)
Stopped Using PWMs	Users have stopped using the PWM. This was only coded for the current, not initial strategy.	- (-%)	4 (1.43%)

mon, such as security increases (46, 16.49%), wanting to add or exclude specific accounts (43, 15.41%), or having problems remembering all accounts or passwords or not wanting to remember them (37, 13.26%). When regarding specific strategies, we found that imports were more often described as convenient, and that adding all accounts at once was more often done out of completion, but less often out of convenience or efficiency. The distribution of reasons per chosen initial strategy is shown in Figure 3.

Finally, we used participant answers to investigate whether they utilized the best-case strategy for security to not only add all of their accounts, but also update every password to a stronger alternative. We found that almost no participants (14, 5.02%) used this approach. Participants mainly argued that adding everything would have been too much work (97, 36.6%), but also that they did not trust their PWM enough to add all important passwords (42, 15.85%), which was a reoccurring theme throughout our whole survey.

We also found participants who stated to be unable to recollect all their accounts (40, 15.09%), as one participant explains: "*I can't even remember that they exist, I can't just suddenly remember all of them and add them to the manager.*" (P142) Other reasons focused on the lack of password changes, including 24 (9.06%) that claimed their passwords were already good enough, or 16 (6.04%) that wanted to keep them, e. g., because they were easily memorizable.

Users were largely happy, but workflows were not seamless: Besides the strategies users applied to initially add their credentials, we were also interested in how they rated their experience, i. e., if they were content with their approach, or encountered any issues that should be mitigated. We therefore asked them both what they liked and went well, and in which situations they struggled with their chosen strategy. We found that in general, a majority of users stated to be satisfied with their experience (157, 56.27%). Some mentioned specific properties they praised, such as PWM features and their usefulness (e. g., browser integration and autosave), that they did not need to remember their passwords anymore, the general increase in security, and how comfortable the whole process was (10.04–12.9%) "*My strategy always worked well, I had to do basically nothing, when the program asked me to add an account I just said yes.*" (P61) However, users also reported negative experiences. Most common were malfunctions or abnormal behavior of both PWMs (29, 10.39%) and websites (25, 8.96%), such as auto-saves not working properly, or too complex password policies. Other mentions included password resets that became necessary due to PWM usage, a lack of overview about their accounts, the high effort and issues due to a lack of synchronization support (3.23–5.02%), for example, one participant detailed why adding accounts on access did not always work for them: "*The accounts which I barely used my strategy didn't work cause I'd end up having to create new passwords time and time again*" (P101)

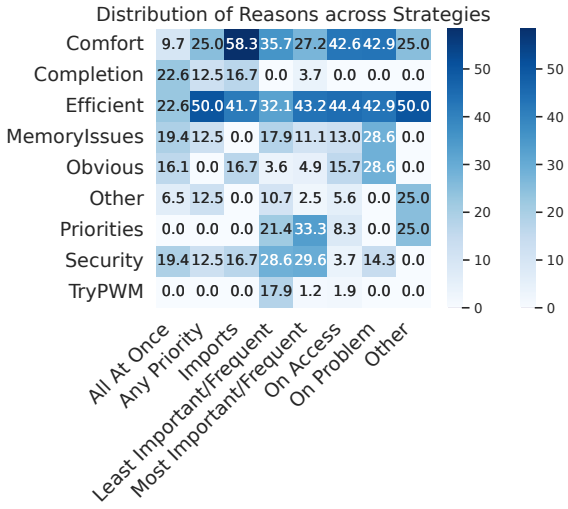


Figure 3: Distribution of reasons for different strategies applied by our participants in %. Our full codebook is described in Appendix D.

We found that despite some negative experiences, the majority of participants (231, 82.8%) stayed loyal to their initial PWM choice. For the 31 (11.11%) participants that switched to a different PWM, common reasons included changing personal devices, e. g., switching to a different browser and therefore using a new built-in PWM (11, 25.58%), and increasing costs and restrictions of free account plans (10, 23.26%).

In addition to asking for situations in which both PWM

usage and the chosen strategy did not work seamlessly, we were interested in participants' wishes and preferences for improved or new features. We found 94 (33.69%) participants who stated to be happy with their PWM, and that they do not need any improvements. Others, who suggested changes, most frequently mentioned the addition of more features, such as better syncing between devices, more memorable password suggestions, or easier bulk imports, (43, 15.41%), or the addition or improvement of automation features (42, 15.05%) such as auto-save, autofill or automated password changes.

This was followed by ways for the PWM to detect their existing accounts based on browser sessions and history, or email inboxes (38, 13.62%), often accompanied by the wish to instantly add these accounts after detection, or improve import features overall (28, 10.04%). However, similar to previous questions, we again found a certain distrust towards PWMs, as 16 (5.73%) participants voiced concerns about their tools collecting too much data, or automated features without them clearly knowing what was happening and why.

5.5.3 General PWM Usage

Besides their initial strategies to set up PWMs, we also asked participants about their general usage to get a broader picture. We found that half of the participants stated to store all their private passwords (131, 46.95%), and an even higher portion that indicated to store every work-related password (183, 65.59%). For those who do not store all private passwords, common reasons included distrust towards the PWM (78, 52.7%) as well as prioritization which passwords should be added (68, 45.95%). We assume that participants are more likely to store all work passwords because they have a lower, more manageable amount that they on average use more frequently.

When asked which website types they prioritized to (not) add, users most commonly mentioned Social Media (78, 27.96%), followed by finance-related websites (47, 16.85%) and email credentials (44, 15.77%).

Since the opportunity to upgrade passwords is an important part of PWM usage, we wanted to know if users did this when they added credentials. Most frequently, they state to have changed some passwords (116, 41.58%), with only 59 (21.15%) who upgraded all of them, and 53 (19.0%) who kept all passwords. We found that reasons to change the passwords were mostly focused around general security increases (35, 22.29%), or to improve weak (55, 35.03%), reused (31, 19.75%) or leaked (14, 8.92%) passwords. Users who decided to keep their old passwords mentioned that their passwords were already strong enough (24, 28.92%), or, a lack of motivation and simply having no reason (26, 31.33%).

> "*And it is not enough to generate a password in the application, you also have to identify yourself on the website, change the password, verify the change... Would you give the same priority to the bank and a game page? Because I do not.*" - P117

When updating existing or new passwords, users typically used the built-in password generators their PWMs offer (56.99–66.86%) or generated the passwords manually (44.0–55.2%).

5.5.4 Comparing Built-in and Third-party PWMs

By design, built-in and third-party PWMs are different in many ways: Their availability, their feature range, as well as how seamless they embed themselves into a users' workflow. Built-in PWMs are by default part of almost any modern browser and operating system. While the feature range and baseline security of built-in PWMs is limited [38, 75], third-party PWMs require a deliberate choice to use the tool, as well as manual installation and some effort to get to know all relevant features. In our survey, we questioned users of both PWM types, and were therefore interested in possible differences between both groups. First, we found that the initial strategies were significantly different ($\chi^2 = 36.04$, $p < 0.005$) between the groups, as third-party users tended to add accounts more often based on their relevance or to add everything at once. However, when regarding the current strategy, the difference was not significant anymore. This is likely because the process of initially adding existing accounts has concluded, and most strategies have shifted towards adding new ones on access. Furthermore, we found significant differences within the reasons for using the respective chosen credential management strategy ($\chi^2 = 20.03$, $p < 0.05$). While built-in users were more often driven by comfort and efficiency, third-party users chose their strategy based on security concerns, and to make sure that either all accounts, or specific important ones were included within their PWM. Similar to this, when asked about their main reason to adopt a PWM ($\chi^2 = 42.85$, $p < 0.005$), we found built-in users more likely to worry about issues such as an overwhelming amount of accounts or forgetting their passwords. However, third-party users were more interested in increasing their overall security, enjoyed the reduced cognitive load of having to memorize only one master password, and were curious to test the PWM, which makes sense as third-party tools require a conscious choice and installation. Additionally, third-party PWM users were significantly more likely to have changed their PWM at some point ($\chi^2 = 26.76$, $p < 0.005$).

6 Ethics & Limitations

In this section, we discuss the ethical considerations and limitations of our work.

Ethics. This work was approved by our institution's Ethical Review Board. We did not collect any personally identifiable information (PII) except anonymous Prolific worker IDs. We

made sure that our survey platform did not collect PII such as IP addresses, and we stored all data on our secured servers respecting GDPR requirements. Only researchers involved in this project had access to the collected information. Before starting our survey, all participants had to sign a consent form detailing the nature and content of our survey, as well as contact information. The consent form also informed all participants that they could quit the survey at any time without repercussions. Finally, we paid all Prolific screening participants $0.37 for their participation in the screening survey, independent of their eligibility for our surveys, and $3.71 for the full survey. With our generous estimates of two minutes for the screening and 20 for the full survey, we paid at least $11 per hour and are in line with Prolific's suggestions for minimum wage survey payment.

Limitations. As is typical for survey studies, our work is affected by self-report bias, recall bias, and social desirability bias. Especially for people whose adoption of the PWM dates further back, these memories may be skewed. Additionally, due to the nature of crowdsourcing platforms such as Prolific, there is also a certain self-selection bias as participants can pick studies they are interested in. However, we decided to use surveys to be able to collect self-reported experiences and thoughts of PWM users that we could not gather using more technical sources such as telemetry or lab studies. In our data analysis, we did not partition participants by the specific PWM they used to get a broader picture of credential management using PWMs. However, it is possible that individual PWMs strongly influence the adoption experience and the participants' satisfaction by, e. g., the presence or absence of certain features. During our early piloting (cf. Section 5.1), we might have missed user strategies. However, as we performed multiple rounds of piloting and collected answers until no new options emerged, we are confident to have gathered all options.

7 Discussion

In this section, we first provide answers to our research questions, then discuss our findings and make recommendations for PWM developers to better help users during PWM setup.

RQ1: *What setup features do password managers offer to new users, who want to add their existing credentials?* Within our expert review (cf. Section 4), we identified setup features present in 14 popular PWMs. These include tutorials and next steps to educate users about the PWMs functions, bulk imports, auto-saves, and account suggestions to help them fill their PWMs. Additionally, many PWMs offer security centers to inform users about their passwords' vulnerability. However, we find these centers often limited to premium versions, especially when regarding breach information. Overall, we find that no feature is available on every PWM and that they are most commonly able to bulk import passwords from other sources or to auto-save them while browsing.

RQ2: *What are common user strategies to add new and existing credentials? Why are these strategies used?* We identified seven user strategies that we mainly differentiate by the time passwords are added (e. g., immediately on install, or when services are accessed) or the choice of which passwords are added (e. g., for more or less important or frequently used accounts), We find the most common strategy to be adding passwords on access, followed by prioritizing more important or frequently used passwords. While present within our sample, other strategies are increasingly irrelevant with time, as users shift more towards adding accounts immediately on creation or when accessing the websites (cf. Section 5.5.2). Overall, users are mostly motivated by convenience or efficiency, and less commonly by security.

RQ3: *How can password manager developers help users with setup, and improve the overall process?* In the following sections, we first discuss our findings from both existing setup features (cf. Section 4), and the credential management strategies and obstacles users report (cf. Section 5). Afterward, we comprise several recommendations for PWM developers, as well as website maintainers.

7.1 Convenience Trumps Security

While researchers generally regard PWMs as security tools, as they enable users to store large amounts of passwords securely and issue complex, randomly generated passwords, end users mostly regard them as convenience tools. We find the main motivation for user strategies is convenience or efficiency, which are named much more commonly than security, confirming previous work on the subject [4, 21]. We additionally find users to voice a certain indifference to accounts they perceive as less important, often because these accounts do not include sensitive information, but also whenever users are not certain whether they will access the account again. This is reflected within the primarily chosen credential management strategies to either add every website whenever it is accessed for the first time, or cherry-pick which accounts are important or frequently used, and skip the storage of others.

7.2 Severe Distrust towards PWMs

In various questions, we find a severe distrust towards PWMs. While research previously found that a lack of trust can decrease PWM adoption [7, 8, 42], our finding is likely also related to recent data breaches and leaks with both LastPass [29] and Norton [9]. Hence, we find negative sentiments not only regarding malicious third parties, but also against PWM vendors that fail to secure user data or are in rare cases even suspected to be the actor that steals data: "*It would be simple for the owner of the software to see that I store most of my things in the application which will make it much more easier for them to hack me.*" (P73) While this does not apply to

every PWM, their security practices, such as used encryption algorithms and key generation settings, are often neither accessible, nor easy to understand for end users. This becomes especially apparent as LastPass has been accused of using weak and insecure security mechanisms after the recent leaks, such as not enforcing long master passwords or not increasing the number of key derivation iterations for older accounts that were created with less secure encryption algorithms [45].

7.3 Complex Account Landscapes

A major issue that may influence users' decisions (not) to add all their accounts at once is that their online landscape can be vast and includes hundreds of more or less important accounts [34, 69]. In our survey, users reported struggling with remembering all accounts and are therefore unable to add them, and begin to prioritize which accounts they insert into their PWM. Furthermore, our survey confirms that adding accounts manually is a tedious and time-consuming process, leading to users choosing more convenient but less secure management strategies, skipping accounts or keeping insecure passwords, therefore not fully utilizing the benefits that come with the usage of a PWM [40, 50].

7.4 Recommendations

Based on our findings, we offer several suggestions for PWM developers and website maintainers to better support end users in the PWM setup process.

Automation: We find that users are motivated by convenience and efficiency, and less security, and therefore argue that processes to add and update passwords should work automated and seamless. Therefore, more automation is necessary. We propose the more widespread use of novel approaches such as well-known URLs for password change [72] that enables PWMs to quickly and easily upgrade passwords without much burden on the user sides. We further suggest creating a standardized format for password imports, to ease the migration between tools.

Account Scans: Besides the tedious task of adding accounts, users are often overwhelmed by their number of accounts, struggle with recollecting them all, and lack reasons to add lesser relevant or used accounts. By using automated scans PWMs can compile account lists by parsing, e. g., registration emails [64] or visited websites. By running these scans locally, they can be designed in a privacy-preserving manner. Based on a participant suggestion, this could be extended to analogue data, such as handwritten notes, by testing the use of optic character recognition, however, future work is required to evaluate its reliability.

Account Suggestions: If automated scans are not feasible, PWMs can suggest either popular sites, or use already added accounts to infer what users might additionally be interested in or whether typical accounts are missing, e. g., suggest adding an email account if none is present in the password database. Furthermore, this could flag, e. g., incomplete or outdated entries that can be deleted.

Guided Additions: We often find imports to require different formats, and especially to require first-time users to compile their own .csv files. To avoid creation and storage of local password lists, we suggest that PWMs offer their own table interface. Using data from either scans and suggestions, or allow users to edit password collections within the encrypted environment, allowing them to collect, revise and quickly add passwords in bulk.

Privacy Labels: We find severe distrust against PWMs, often due to unclear encryption and security mechanisms. Previous work presented privacy labels that both deliver information regarding the incorporated security, but also allow non-experts to quickly assess the vulnerability of their product [19, 37]. We argue that this could be helpful to address distrust that stems from a lack of knowledge and familiarity, and suggest that PWM developers adopt privacy labels for their programs.

Gamification & Nudges: We find that some users mention that adding accounts is a tedious task, and find one PWM to offer achievements for users to both introduce the PWMs main functions and motivate them to actively fill it. Since gamification has shown promising results in other areas [25, 56], we argue that it should be evaluated for PWMs as well. Related, PWMs could add nudges to motivate and remind users more firmly to store or update passwords at risk [6, 22].

8 Conclusion

The main benefits from using PWMs stems from adding all accounts as soon as possible, and updating every password to a strong alternative, since there is no need to memorize them anymore. In this work, we identified common setup features within 14 popular PWMs, and surveyed 279 end users regarding their credential management strategies. We find that while PWMs offer various setup features that help users add credentials and set secure passwords such as imports, autosave functions or password scoring, they are not present in all PWMs. We identified seven strategies users apply during PWM setup, most commonly adding passwords when accessing websites or when they are perceived as important. We found that end users are mainly motivated by convenience and efficiency, not password security. However, we also noticed distrust towards PWMs, leading to users not storing their accounts as they are concerned for the safety of their data. Due to a lack of motivation, perceived necessity and distrust towards PWMs, they often refrain from adding all accounts, thereby severely limiting the gained security benefits from using a PWM. Finally, we propose several recommendations how this problem can be approached by PWM developers, including more automated workflows and methods to increase user motivation.

Acknowledgements

We thank all participants for their valuable time and insights shared with us. Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC 2092 CASA – 390781972.

References

[1] 1Password. The World's Most-Loved Password Manager. https://1password.com/ (visited on 01/18/2023).

[2] Ruba Abu-Salma, Elissa M. Redmiles, Blase Ur, and Miranda Wei. Exploring User Mental Models of End-to-End Encrypted Communication Tools. In *Proc. 8th USENIX Workshop on Free and Open Communications on the Internet (FOCI'18)*. USENIX Association, 2018.

[3] Yusuf Albayram, John Liu, and Stivi Cangonj. Comparing the Effectiveness of Text-based and Video-based Delivery in Motivating Users to Adopt a Password Manager. In *Proceedings of the 2021 European Symposium on Usable Security*, pages 89–104, 2021.

[4] Nora Alkaldi and Karen Renaud. Why Do People Adopt, or Reject, Smartphone Password Managers? In *1st European Workshop on Usable Security*. Internet Society, 2016.

[5] Nora Alkaldi and Karen Renaud. Encouraging Password Manager Adoption by Meeting Adopter Self-Determination Needs. In *Proceedings of the 52nd Hawaii International Conference on System Sciences*, 2019.

[6] Hazim Almuhimedi, Florian Schaub, Norman Sadeh, Idris Adjerid, Alessandro Acquisti, Joshua Gluck, Lorrie Faith Cranor, and Yuvraj Agarwal. Your Location Has Been Shared 5,398 Times! A Field Study on Mobile App Privacy Nudging. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 787–796, 2015.

[7] Salvatore Aurigemma, Thomas Mattson, and Lori Leonard. So Much Promise, So Little Use: What is Stopping Home End-Users from Using Password Manager Applications? In *Proceedings of the 50th Hawaii International Conference on System Sciences*, 2017.

[8] Ramakrishna Ayyagari, Jaejoo Lim, and Olger Hoxha. Why Do Not We Use Password Managers? A Study on the Intention to Use Password Managers. *Contemporary Management Research*, 15(4):227–245, 2019.

[9] Bill Toulas. NortonLifeLock Warns That Hackers Breached Password Manager Accounts. https://www.bleepingcomputer.com/news/security/nortonlifelock-warns-that-hackers-breached-password-manager-accounts/ (visited on 2/14/2023).

[10] Bitwarden. Auto save newly created login info. https://community.bitwarden.com/t/auto-save-newly-created-login-info/13555/28 (visited on 05/18/2023).

[11] Joseph Bonneau, Cormac Herley, Paul C van Oorschot, and Frank Stajano. The Quest to Replace Passwords: A Framework for Comparative Evaluation of Web Authentication Schemes. In *Proc. 33rd IEEE Symposium on Security and Privacy (SP'12)*. IEEE, 2012.

[12] Jason Ceci, Hassan Khan, Urs Hengartner, and Daniel Vogel. Concerned but Ineffective: User Perceptions, Methods, and Challenges when Sanitizing Old Devices for Disposal. In *SOUPS @ USENIX Security Symposium*, pages 455–474, 2021.

[13] Sunil Chaudhary, Tiina Schafeitel-Tähtinen, Marko Helenius, and Eleni Berki. Usability, Security and Trust in Password Managers: A Quest for User-Centric Properties and Features. *Computer Science Review*, 33:69–90, 2019.

[14] Sonia Chiasson and Paul C van Oorschot. Quantifying the Security Advantage of Password Expiration Policies. *Designs, Codes and Cryptography*, 77(2):401–408, 2015.

[15] Sonia Chiasson, Paul C van Oorschot, and Robert Biddle. A Usability Study and Critique of Two Password Managers. In *USENIX Security Symposium*, volume 15, pages 1–16, 2006.

[16] Victoria Clarke and Virginia Braun. *Thematic Analysis*, pages 1947–1952. Springer New York, New York, NY, 2014.

[17] James S Conners and Daniel Zappala. Let's Authenticate: Automated Cryptographic Authentication for the Web with Simple Account Recovery. *Who Are You*, 2019.

[18] Dashlane. Password Manager App for Home, Mobile, Business. https://www.dashlane.com/ (visited on 01/18/2023).

[19] Pardis Emami-Naeini, Janarth Dheenadhayalan, Yuvraj Agarwal, and Lorrie Faith Cranor. Which Privacy and Security Attributes Most Impact Consumers' Risk Perception and Willingness to Purchase IoT Devices? In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 519–536. IEEE, 2021.

[20] Shayan Eshkandary, David Barrera, Elizabeth Stobert, and Jeremy Clark. A First Look at the Usability of Bitcoin Key Management. In *NDSS Symposium 2015*, page 05_3_3. Internet Society, 2015.

[21] Michael Fagan, Yusuf Albayram, Mohammad Maifi Hasan Khan, and Ross Buck. An Investigation Into Users' Considerations Towards Using Password Managers. *Human-centric Computing and Information Sciences*, 7(1):1–20, 2017.

[22] Felix Fischer, Huang Xiao, Ching-Yu Kao, Yannick Stachelscheid, Benjamin Johnson, Danial Razar, Paul Fawkesley, Nat Buckley, Konstantin Böttinger, Paul Muntean, and Jens Grossklags. Stack Overflow Considered Helpful! Deep Learning Security Nudges Towards Stronger Cryptography. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 339–356, Santa Clara, CA, August 2019. USENIX Association.

[23] Xianyi Gao, Yulong Yang, Can Liu, Christos Mitropoulos, Janne Lindqvist, and Antti Oulasvirta. Forgetting of Passwords: Ecological Theory and Data. In *27th USENIX Security Symposium (USENIX Security 18)*, pages 221–238, 2018.

[24] Hana Habib, Jessica Colnago, William Melicher, Blase Ur, Sean Segreti, Lujo Bauer, Nicolas Christin, and Lorrie Cranor. Password Creation in the Presence of Blacklists. *Proc. USEC*, page 50, 2017.

[25] Katrin Hartwig, Atlas Englisch, Jan Pelle Thomson, and Christian Reuter. Finding Secret Treasure? Improving Memorized Secrets Through Gamification. In *Proceedings of the 2021 European Symposium on Usable Security*, pages 105–117, 2021.

[26] Nicolas Huaman, Sabrina Amft, Marten Oltrogge, Yasemin Acar, and Sascha Fahl. They Would do Better if They Worked Together: The Case of Interaction Problems Between Password Managers and Websites. In *Proc. 42nd IEEE Symposium on Security and Privacy (SP'21)*. IEEE, 2021.

[27] Troy Hunt. Have I Been Pwned: Check If Your Email Has Been Compromised in a Data Breach. https://haveibeenpwned.com/ (visited on 10/20/2021).

[28] Philip G Inglesant and M Angela Sasse. The True Cost of Unusable Password Policies: Password Use in the Wild. In *Proceedings of the sigchi conference on human factors in computing systems*, pages 383–392. ACM, 2010.

[29] Karim Toubba. Notice of Recent Security Incident. https://blog.lastpass.com/2022/12/notice-of-recent-security-incident/ (visited on 2/14/2023).

[30] Ambarish Karole, Nitesh Saxena, and Nicolas Christin. A Comparative Usability Evaluation of Traditional Password Managers. In *International Conference on Information Security and Cryptology*, pages 233–251. Springer, 2010.

[31] Harjot Kaur, Sabrina Amft, Daniel Votipka, Yasemin Acar, and Sascha Fahl. Where to Recruit for Security Development Studies: Comparing Six Software Developer Samples. In *31st USENIX Security Symposium, USENIX Security '22, Boston MA, USA, August 10-12, 2022*. USENIX Association, Aug 2022.

[32] KeePassXC. KeePassXC - Cross-Platform Password Manager. https://keepassxc.org/ (visited on 01/18/2023).

[33] Keeper. Generate a Strong Random Password. https://www.keepersecurity.com/password-generator.html (visited on 09/09/2021).

[34] Limor Kessem. Surge of New Digital Accounts During the Pandemic Leads to Lingering Security Side Effects. https://securityintelligence.com/posts/new-digital-accounts-pandemic-security-side-effects/ (visited on 01/18/2023).

[35] LastPass. LastPass | Password Manager. https://www.lastpass.com/ (visited on 01/18/2023).

[36] Kevin Lee, Sten Sjöberg, and Arvind Narayanan. Password Policies of Most Top Websites Fail to Follow Best Practices. In *Eighteenth Symposium on Usable Privacy and Security (SOUPS 2022)*, pages 561–580, 2022.

[37] Yucheng Li, Deyuan Chen, Tianshi Li, Yuvraj Agarwal, Lorrie Faith Cranor, and Jason I Hong. Understanding iOS Privacy Nutrition Labels: An Exploratory Large-Scale Analysis of App Store Data. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, pages 1–7, 2022.

[38] Zhiwei Li, Warren He, Devdatta Akhawe, and Dawn Song. The Emperor's New Password Manager: Security Analysis of Web-Based Password Managers. In *23rd USENIX Security Symposium (USENIX Security 14)*, pages 465–479, 2014.

[39] Gabriel Lima, Nina Grgić-Hlača, and Meeyoung Cha. Human Perceptions on Moral Responsibility of AI: A Case Study in AI-assisted Bail Decision-Making. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2021.

[40] Sanam Ghorbani Lyastani, Michael Schilling, Sascha Fahl, Michael Backes, and Sven Bugiel. Better Managed than Memorized? Studying the Impact of Managers on

Password Strength and Reuse. In *27th USENIX Security Symposium (USENIX Security 18)*, pages 203–220, 2018.

[41] Sanam Ghorbani Lyastani, Michael Schilling, Michaela Neumayr, Michael Backes, and Sven Bugiel. Is FIDO2 the Kingslayer of User Authentication? A Comparative Usability Study of FIDO2 Passwordless Authentication. In *2020 IEEE Symposium on Security and Privacy (SP)*, pages 268–285. IEEE, 2020.

[42] Raymond Maclean and Jacques Ophoff. Determining Key Factors that Lead to the Adoption of Password Managers. In *2018 International Conference on Intelligent and Innovative Computing Applications (ICONIC)*, pages 1–7. IEEE, 2018.

[43] Peter Mayer, Collins W Munyendo, Michelle L Mazurek, and Adam J Aviv. Why Users (Don't) Use Password Managers at a Large Educational Institution. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 1849–1866, 2022.

[44] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. Reliability and Inter-Rater Reliability in Qualitative Research: Norms and Guidelines for CSCW and HCI Practice. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–23, 2019.

[45] Mitchell Clark. The LastPass Disclosure of Leaked Password Vaults is Being Torn Apart by Security Experts. https://www.theverge.com/2022/12/28/23529547/lastpass-vault-breach-disclosure-encryption-cybersecurity-rebuttal (visited on 2/14/2023).

[46] NordPass. How Secure is my Password? https://nordpass.com/secure-password/ (visited on 01/18/2023).

[47] Sean Oesch, Scott Ruoti, James Simmons, and Anuj Gautam. "It Basically Started Using Me:" An Observational Study of Password Manager Usage. In *CHI Conference on Human Factors in Computing Systems*, pages 1–23, 2022.

[48] Stefan Palan and Christian Schitter. Prolific.ac — A Subject Pool for Online Experiments. *Journal of Behavioral and Experimental Finance*, 17:22–27, 2018.

[49] Sarah Pearman, Jeremy Thomas, Pardis Emami Naeini, Hana Habib, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, Serge Egelman, and Alain Forget. Let's Go in for a Closer Look: Observing Passwords in Their Natural Habitat. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 295–310, 2017.

[50] Sarah Pearman, Shikun Aerin Zhang, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. Why People (Don't) Use Password Managers Effectively. In *Fifteenth Symposium On Usable Privacy and Security (SOUPS 2019). USENIX Association, Santa Clara, CA*, pages 319–338, 2019.

[51] Katharina Pfeffer, Alexandra Mai, Adrian Dabrowski, Matthias Gusenbauer, Philipp Schindler, Edgar Weippl, Michael Franz, and Katharina Krombholz. On the Usability of Authenticity Checks for Hardware Security Tokens. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 37–54, 2021.

[52] Denise Ranghetti Pilar, Antonio Jaeger, Carlos FA Gomes, and Lilian Milnitsky Stein. Passwords Usage and Human Memory Limitations: A Survey Across Age and Educational Background. *PloS one*, 7(12):e51067, 2012.

[53] Peter G Polson, Clayton Lewis, John Rieman, and Cathleen Wharton. Cognitive Walkthroughs: A Method for Theory-based Evaluation of User Interfaces. *International Journal of man-machine studies*, 36(5):741–773, 1992.

[54] Stanley Presser, Mick P Couper, Judith T Lessler, Elizabeth Martin, Jean Martin, Jennifer M Rothgeb, and Eleanor Singer. Methods for Testing and Evaluating Survey Questions. *Methods for Testing and Evaluating Survey Questionnaires*, pages 1–22, 2004.

[55] Prolific. Prolific | Online Participant Recruitment for Surveys and Market Research. https://prolific.co/, 2020.

[56] George E Raptis, Christina Katsini, Andrew Jian-Lan Cen, Nalin Asanka Gamagedara Arachchilage, and Lennart E Nacke. Better, Funner, Stronger: A Gameful Approach to Nudge People into Making Less Predictable Graphical Password Choices. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2021.

[57] Hirak Ray, Flynn Wolf, Ravi Kuber, and Adam J Aviv. Why Older Adults (Don't) Use Password Managers. *arXiv preprint arXiv:2010.01973*, 2020.

[58] Joel Reardon, David Basin, and Srdjan Capkun. SoK: Secure Data Deletion. In *2013 IEEE Symposium on Security and Privacy*, pages 301–315. IEEE, 2013.

[59] Rebecca Stone. LastPass Now Offers Time-Based One-Time Passcode (TOTP). https://blog.lastpass.com/2020/12/lastpass-now-offers-time-based-one-time-passcode-totp/ (visited on 1/18/2023).

[60] Elissa M Redmiles, Yasemin Acar, Sascha Fahl, and Michelle L Mazurek. A Summary of Survey Methodology Best Practices for Security and Privacy Researchers. Technical report, 2017.

[61] John Rieman, Marita Franzke, and David Redmiles. Usability Evaluation with the Cognitive Walkthrough. In *Conference companion on Human factors in computing systems*, pages 387–388, 1995.

[62] Daniel Russo and Klaas-Jan Stol. Gender Differences in Personality Traits of Software Engineers. *IEEE Transactions on Software Engineering*, 2020.

[63] Saferpass. Credit Card Support. https://saferpass.net/credit-cards (visited on 1/18/2023).

[64] Paul Sawers. Dashlane Launches Mobile Email Inbox Scanning to Assess Your Online Security Hygiene, 2018. https://venturebeat.com/2018/06/27/dashlane-launches-mobile-email-inbox-scanning-to-assess-your-online-security-hygiene/ (visited on 01/18/2023).

[65] Sunyoung Seiler-Hwang, Patricia Arias-Cabarcos, Andrés Marín, Florina Almenares, Daniel Díaz-Sánchez, and Christian Becker. "I Don't See Why I Would Ever Want to Use It" Analyzing the Usability of Popular Smartphone Password Managers. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pages 1937–1953, 2019.

[66] James Simmons, Oumar Diallo, Sean Oesch, and Scott Ruoti. Systematization of password manageruse cases and design paradigms. In *Annual Computer Security Applications Conference*, pages 528–540, 2021.

[67] Frank Stajano. Pico: No More Passwords! In *Security Protocols XIX - 19th International Workshop*. Springer, 2011.

[68] statcounter GlobalStats. Chrome Market Share. https://gs.statcounter.com/browser-market-share (visited on 01/18/2023).

[69] Jack Turner. Study Reveals Average Person Has 100 Passwords. https://tech.co/news/average-person-100-passwords (visited on 06/23/2021).

[70] Blase Ur, Jonathan Bees, Sean M Segreti, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. Do Users' Perceptions of Password Security Match Reality? In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 3748–3760, 2016.

[71] Blase Ur, Fumiko Noma, Jonathan Bees, Sean M Segreti, Richard Shay, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. "I Added '!'at the End to Make It Secure": Observing Password Creation in the Lab. In *Symposium on Usable Privacy and Security (SOUPS)*, 2015.

[72] W3C. A Well-Known URL for Changing Passwords. https://w3c.github.io/webappsec-change-password-url/ (visited on 11/29/2022).

[73] Rick Wash, Emilee Rader, Ruthie Berman, and Zac Wellmer. Understanding Password Choices: How Frequently Entered Passwords are Re-used Across Websites. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*, pages 175–188, 2016.

[74] Dominik Wermke, Christian Stransky, Nicolas Huaman, Niklas Busch, Yasemin Acar, and Sascha Fahl. Cloudy with a Chance of Misconceptions: Exploring Users' Perceptions and Expectations of Security and Privacy in Cloud Office Suites. In *Sixteenth Symposium on Usable Privacy and Security, SOUPS 2020, August 12-14, 2020*, Aug 2020.

[75] Rui Zhao and Chuan Yue. All Your Browser-Saved Passwords Could Belong to Us: A Security Analysis and a Cloud-Based New Design. In *Proceedings of the third ACM conference on Data and application security and privacy*, pages 333–340, 2013.

[76] Samira Zibaei, Dinah Rinoa Malapaya, Benjamin Mercier, Amirali Salehi-Abari, and Julie Thorpe. Do Password Managers Nudge Secure (Random) Passwords? In *Eighteenth Symposium on Usable Privacy and Security (SOUPS 2022)*, pages 581–597, Boston, MA, August 2022. USENIX Association.

A Appendix: Expert Review Tasks

1.1 Install PWM:

- (Omit and proceed with 1.4 if only a browser extension is found)
- Go to the respective website
- Find, download and install the PWM
- Start full-desktop recording
- Choose the most basic plan available (usually free or premium-trial)

1.2 Set a Bad Master Password:

- When prompted for the master password, set a bad password
- If required, provide a secure alternative

1.3 Search for Setup Features

- Follow the setup flow provided by the PWM
- Follow all references to tutorials, next steps, additional information, getting started areas, and similar

1.4 Install the Browser Extension:

- If available, install the complementary browser extension of the PWM
- If the PWM has no standalone version: Start full-desktop recording

- If the PWM has no standalone version: Perform Steps 1.2 and 1.3 for the browser extension

2.1 Search for a Mass Import Feature

- Search the PWM for a mass import feature
- If you cannot find the feature, conduct a Google search: "<PWM name> import feature"

2.2 Add a Password for a Popular Website

- Create a new account entry and add a password for a very popular website.
- Delete entry afterward

2.3 Add a Password for a Less Popular Website

- Create a new account entry and add a password for a less popular website

2.4 Generate a Password for Any Website

- Create a new entry and *generate* a password for any website

2.5 Add a Bad Password for Any Website

- Create a new entry and add a bad password for any website

2.6 Add a Reused Password for any Website

- Create a new entry and reuse a strong password for any website

2.7 Add the Master Password for Any Website

- Open any entry and add the current master password as password

2.8 Add a Password via Autosave

- Open a website and try to log in with activated *autosave* functions (if available)

3.1 Search for a Security Center

- Search the PWM for a security center
- If you cannot find a security center, conduct a Google search: [<PWM name> security center]

3.2 Check Password Strength

- Search for a feature to rate password strength (especially password meters)

3.3 Check Passwords for Leaks/Breaches

- Search for a feature to check passwords and other data for their appearances in leaks

4 Add Two-Factor Authentication

- Open any entry and add a (predetermined) two-factor authentication seed to it

B Appendix: Screening survey

SQ1 How do you manage your passwords? [multiple choice]

- ☐ I memorize them
- ☐ I write them on a piece of paper
- ☐ I keep them in a (hidden) textfile
- ☐ My browser/phone remembers them for me
- ☐ I am using a password manager
- ☐ Other (please specify): [free text]

SQ2 [If "My browser/phone remembers them for me" or "I am using a password manager" in SQ1:] When did you start using a password manager? [single choice]

- ○ In the last week
- ○ In the last month, but not in the last week
- ○ In the last six months, but not in the last month
- ○ In the last two years, but not in the last six months
- ○ More than two years ago
- ○ Never, and I do not plan to
- ○ Never, but I do plan to
- ○ I do not know / I do not remember

SQ3 [If "My browser/phone remembers them for me" or "I am using a password manager" in SQ1:] Please name your first password manager(s). [free text]

SQ4 [If "My browser/phone remembers them for me" or "I am using a password manager" in SQ1:] On which devices are you/have you been using a Password Manager? [multiple choice]

- ☐ Windows
- ☐ Linux
- ☐ Mac
- ☐ iOS
- ☐ Android
- ☐ Blackberry
- ☐ ChromeOS
- ☐ Other (please specify): [free text]

SQ5 [If neither "My browser/phone remembers them for me" nor "I am using a password manager" in SQ1:] Why are you not using a password manager? [free text]

C Appendix: Survey

Password managers are tools that can help you store passwords and create strong ones. This includes both programs or browser extensions you chose and installed (e. g., LastPass, 1Password), and the password managers included in your phone or browser (e. g., Chrome, Apple Keychain).

Q1 When did you start using a password manager? [single choice]

- ○ In the last week
- ○ In the last month, but not in the last week
- ○ In the last six months, but not in the last month
- ○ In the last two years, but not in the last six months
- ○ More than two years ago
- ○ Never, and I do not plan to
- ○ Never, but I do plan to
- ○ I do not know / I do not remember

Q2 Please name your first password manager(s) [free text]

Q3 Are you still using the same password manager? If not, which one are you currently using, and why did you decide to change? [free text]

Q4 Are you paying for your password manager? [single choice]

- ○ Yes
- ○ No

Q5 How likely is it that you would recommend a password manager to a friend or colleague? [Net Promoter Score, 11-point likert from *Not at all likely* to *Extremely likely*]

Q6 Are all of your private accounts stored inside of a password manager? [single choice]

- ○ Yes, all of them
- ○ No, not all of them
- ○ I am not sure

Q7 [If “No, not all of them” or “I am not sure” in Q6] Please share with us why you are not storing all of your private accounts in a password manager. [free text]

Q8 Are all of your work accounts stored inside of a password manager? [single choice]

- ○ Yes, all of them
- ○ No, not all of them
- ○ I am not sure

Q9 [If “No, not all of them” or “I am not sure” in Q8] Please share with us why you are not storing all of your work accounts in a password manager. [free text]

Q10 Please try to remember your thoughts when you first started using a password manager. Which of the reasons below best describe your main reasons to use a password manager? Please select all that apply to you. [multiple choice]

- ☐ I wanted to increase security (e. g. I could use stronger passwords)
- ☐ Creating passwords myself was tiresome
- ☐ I only needed to remember one master password
- ☐ It became easier to organize my passwords
- ☐ All my passwords were in one place
- ☐ I had too many accounts to remember my passwords
- ☐ I kept forgetting my passwords
- ☐ It was a requirement by my employer
- ☐ I was curious to test password managers
- ☐ Somebody recommended using them to me
- ☐ The password manager can generate strong passwords
- ☐ The free trial convinced me to use it
- ☐ Other (please specify:) [free text]

Q11 [If “Somebody recommended using them to me” in Q10] Who recommended using a password manager to you? Please choose all that apply. [multiple choice]

- ☐ A spouse/significant other
- ☐ Family
- ☐ Friends
- ☐ Colleagues
- ☐ Somebody else (please specify:) [free text]
- ☐ Nobody
- ☐ I do not remember
- ☐ Prefer not to disclose

Q12 Were you or somebody you know the victim of a data breach or hack that leaked all or some of your login information (e. g. passwords, email addresses, hashes)? [single choice]

- ○ Yes
- ○ No
- ○ I do not know
- ○ Prefer not to disclose

Q13 For this question, try to remember how you started out with the password manager. What was your initial strategy to add existing accounts, e. g., did you add them all at once or did you prioritize certain accounts? Please include as many details as you can recall. [free text]

Q14 Which of the strategies below fits your main strategy to add existing accounts best? [single choice]

- ○ Added them whenever the account was accessed or visited
- ○ Started with rarely used accounts
- ○ Started with less important accounts
- ○ Started with more important accounts
- ○ Added all I could remember at once
- ○ Added them whenever I encountered a problem (e. g. needed to reset the password)
- ○ Imported my passwords from e. g. my browser
- ○ Other (please specify:) [free text]
- ○ I do not remember my main strategy

Q15 Why did you use your particular strategy of adding existing accounts to your password manager? (e. g., because it was time-efficient or less work to add only certain accounts, or because it increased security to add all accounts at once) [free text]

Q16 Please share your current strategy to add new or existing accounts into your password manager with us. We are especially interested in how it differs from your initial strategy, and why you decided to change your approach. If you kept your initial strategy, please write “no”. [free text]

Q17 If you stated to prioritize certain accounts in any of the previous questions: Please specify the type, e. g., social media. [free text]

Q18 Did you update your existing passwords when you added accounts to the password manager? [single choice]

- ○ Yes, all of them
- ○ Yes, some of them
- ○ No, none of them
- ○ I do not remember
- ○ Prefer not to disclose

Q19 [If “Yes, all of them” in Q18] Please elaborate in detail why you updated all of your passwords. [free text]

Q20 [If “Yes, some of them” in Q18] Please elaborate in detail why you updated some of your passwords, but not all of them. [free text]

Q21 [If “No, none of them” in Q18] Please elaborate in detail why you updated none of your passwords. [free text]

Q22 [If “Yes, all of them” or “Yes, some of them” in Q18] In which way do you update your **existing passwords** when adding them to your password manager? Please choose all that apply. [multiple choice]

- ☐ Update them using the password manager’s built-in password generator
- ☐ Update them using an external password generator
- ☐ Update them with a manually created new password
- ☐ I do not add existing accounts
- ☐ Other (please specify:) [free text]

Q23 In which way do you generate your passwords when creating new accounts and adding them to your password manager? Please choose all that apply. [multiple choice]

- ☐ Generate them using the password manager’s built-in password generator
- ☐ Generate them using an external password generator
- ☐ Manually create a new password
- ☐ I do not add new accounts
- ☐ Other (please specify:) [free text]

Q24 Please share your experiences with initially adding your passwords to your password manager with us. In which situations and for which accounts did your strategy work well? [free text]

Q25 In which situations and for which accounts did you stumble over problems with your strategy? Which problems did occur? [free text]

Q26 In addition to storing passwords, you use your password manager for: [multiple choice]

- ☐ Autofilling Two-Factor Authentication
- ☐ Storing banking or credit card information
- ☐ Storing address information
- ☐ Storing secret notes (e. g., Recovery codes, SSH keys, private encryption keys)
- ☐ Storing other data (please specify:) [free text]
- ☐ Checking your password strength
- ☐ Checking if your passwords were part of a data breach
- ☐ Checking your passwords for reuse
- ☐ Generating strong passwords
- ☐ I am not using additional functions
- ☐ Other functions (please specify:) [free text]

Q27 [If not "Added all I could remember at once" in Q14 or "Yes, some of them" or "No, none of them" in Q18] You have stated that you did not add and update all passwords at once when you initially started using your password manager. We are interested to learn why you did not do this. Please provide details for your reasoning. [free text]

Q28 In which ways could your password manager have supported **your** process of initially adding all your existing passwords better? Please assume that there are no technical limitations, e. g. that password managers can access all data they need. [free text]

Q29 What is your gender? We use this information to increase visibility of less represented genders. [single choice]

- ○ Woman
- ○ Man
- ○ Non-binary
- ○ Prefer not to disclose
- ○ Prefer to self-describe [free text]

Q30 What is your age in years? [integer input]

Q31 Which of the following describe your race and ethnicity, if any? Please check all that apply. [multiple choice]

- ☐ White or of European descent
- ☐ South Asian
- ☐ Hispanic or Latino/a/x
- ☐ Middle Eastern
- ☐ East Asian
- ☐ Black or of African descent
- ☐ Southeast Asian
- ☐ Indigenous (such as Native American, Pacific Islander, or Indigenous Australian)
- ○ Prefer not to disclose
- ○ Prefer to self-describe [free text]

Q32 Which of the following best describes the highest level of formal education that you have completed? [single choice]

- ☐ I never completed any formal education
- ☐ 10th grade or less (e. g., some American high school credit, German Realschule, British GCSE)
- ☐ Secondary school (e. g., American high school, German Realschule or Gymnasium, Spanish or French Baccalaureate, British A-Levels)
- ☐ Trade, technical or vocational training
- ☐ Some college/university study without earning a degree
- ☐ Associate degree (A.A., A.S., etc.)
- ☐ Bachelor's degree (B.A., B.S., B.Eng., etc.)
- ☐ Master's degree (M.A., M.S., M.Eng., MBA, etc.)
- ☐ Professional degree (JD, MD, etc.)
- ☐ Other doctoral degrees (Ph.D., Ed.D., etc.)
- ○ Prefer not to disclose
- ○ Other (please specify:) [free text]

Q33 Do you have a formal education (Bachelor's degree or higher) in computer science, information technology, or a related field? [single choice]

- ○ Yes
- ○ No
- ○ Prefer not to disclose

Q34 Have you held a job in computer science, information technology, or a related field? [single choice]

- ○ Yes
- ○ No
- ○ Prefer not to disclose

D Appendix: Codebook

Table 4: The codebook with descriptions we used to code the open-ended questions. For questions Q13 and Q16 see Table 3.

Code	Q3	Q7	Q9	Q15	Q17	Q19	Q20	Q21	Q24	Q25	Q27	Q28	Description
Accounts										■	■		Problems gathering accounts, e. g., due to amount of accounts, lack of overview, old or rarely used accounts.
Administrative					■								Prioritization of governmental, civic or medical accounts.
Automatization												■	Desire for better automatization, e. g., autosave, autofill, or autochange of passwords.
Breach						■	■			■			(Some) Passwords appeared in a breach or data leak, passwords were changed to be safe.
ChangedDevice	■												Device, browser, or workplaces have changed, and the old PWM is not available or practical anymore.
Comfort				■		■	■		■				The process of, e. g., adding passwords, changing passwords, improving passwords, is easy to complete, or even enjoyable.
Completion				■									Desire to add all accounts into the PWM as soon as possible.
Cost	■												Old PWM became too expensive, or adjusted the feature range of its account plans (including free plans).
Distrust		■	■								■	■	Skepticism or fear towards the PWM, e. g., because it might break, leak data, or spy on its user, which influences participants behavior or perception.
EaseOfLogin									■				PWM eases login processes or remembers passwords, meaning the user no longer has to.
EaseOfUse	■												New PWM is easier to use and or more comfortable.
Education					■								Prioritization of educational accounts, e. g., school or university accounts, or platforms with educational content.
Efficient				■					■				Process of, e. g., adding or changing passwords, is very fast or simple (in terms of time or effort).
Effort		■	■				■	■		■	■		Strategy or process to, e. g., add or change passwords is too time-consuming, cumbersome, or too much work.
Email					■								Prioritization of email accounts.
Financial					■								Prioritization of accounts related to finances, e. g., banking accounts or cryptocurrency wallets.
GoodPWs							■	■			■		(Some) Passwords are good enough, no need to change them, or the account is already secure enough due to, e. g., multi-factor authentication.
Guidance												■	Desire to have more guides, tips and tricks, or explanation on how the PWM works.
Imports												■	Bulk password imports from different sources, e. g., browser, other PWMs, or .csv files.
HighSecurity					■								Prioritization of accounts with special access rights or containing sensitive or personal data.
MemoryIssues				■		■	■						Issues related to not being able or not wanting to remember accounts or passwords, e. g., it is hard to remember passwords without PWMs.
Misc					■								Prioritization of rarely mentioned account types, or those too unspecific or all-encompassing to assign them to a more specific code, e. g., “Google”.
MultipleDevices		■	■							■			Multiple devices or platforms are used, but the PWM is not (easily) available on all of them, e. g., participant cannot access PWM on their phone or devices such as smart TVs.
Obvious				■									Obvious choice, no good alternative, or most logical strategy.
OpenSource	■												The new PWM is open source and hence more trustworthy.
PasswordReset										■			Participant has to reset passwords they could not remember.
Priorities		■	■	■			■						Prioritization of specific accounts, e. g., (not) adding/updating accounts because they are (not) important, rarely/often used, or include sensitive data. Note that Q17 codes detailed answers regarding which accounts were prioritized.
PWMFeatures	■								■			■	Convenience due to PWM features and functions in all steps of the process, e. g., suggestions of popular accounts, allowing more content fields within password entries, or allowing syncing.
PWMLimitations		■								■			Problems due to PWM properties, e. g., features missing or not working (well) or limitations (e. g., only limited number of accounts or devices possible).
PWScoring												■	Desire for feedback regarding the password strength, and reused or breached passwords, and suggestion of better passwords.
QualityOfLife												■	Desire for better interfaces, easier handling, or in general higher usability of the PWM.
RegularUpdate						■	■						(Some) Passwords are changed based on undefined trigger or external policies.
Remembrance		■	■				■	■		■	■		Participant has easy memorizable passwords, or prefers to keep memorizable passwords.
Reused						■	■						(Some) Password were reused or only slightly modified, passwords were changed to be safe.
Satisfied	■								■	■		■	Participant is satisfied (with current situation) or highlights neither good nor negative aspects.
Scans												■	Desire to receive a list of websites/accounts based on scanning emails, history, or handwriting (OCR).
Security	■			■		■	■		■				Behavior, e. g., adding accounts, changing passwords or PWMs, that helps mitigate security issues or with the goal to increase security overall.
Shopping					■								Prioritization of online shops or accounts related to shopping.
SocialMedia					■								Prioritization of social media or forum accounts.
TryPWM				■							■		Participant (initially) wanted to test the PWM, which influences the strategy.
Unknown													Participant did initially not know about the strategy or did not think about using it.
Unwilling							■	■			■		Participant does not want to try the strategy or thinks it is not necessary, e. g., does not see the need to add/update more than the most important passwords.
WeakPWs						■	■						(Some) Password were weak, too old or considered bad for other reasons, passwords were changed to be safe.
Websites										■			Websites obstructed process, e. g., by adding complex password requirements or disabling autofill.
Work					■								Prioritization of work-related accounts, e. g., work accounts or career websites.
Workplace			■										PWM usage only for work or influenced by work (requirements).

■ Code used at least once for the respective question.

Evolution of Password Expiry in Companies: Measuring the Adoption of Recommendations by the German Federal Office for Information Security

Eva Gerlitz
Fraunhofer FKIE

Maximilian Häring
University of Bonn

Matthew Smith
University of Bonn, Fraunhofer FKIE

Christian Tiefenau
University of Bonn

Abstract

In 2020, the German Federal Office for Information Security (BSI) updated its Password composition policy (PCP) guidelines for companies. This included the removal of password expiry, which research scholars have been discussing for at least 13 years. To analyze how the usage of password expiry in companies evolved, we conducted a study that surveyed German companies three times: eight months ($n = 52$), two years ($n = 63$), and three years ($n = 80$) after these changed recommendations. We compared our results to data gathered shortly before the change in 2019. We recruited participants via the BSI newsletter and found that 45% of the participants said their companies still use password expiry in 2023. The two main arguments were a) to increase security and b) because some stakeholders still required these regular changes. We discuss the given reasons and offer suggestions for research and guiding institutions.

1 Introduction

Password composition policies (PCPs) aim to increase account security. Yet, research has shown that individuals often devise strategies to deal with PCPs to make them less unpleasant, resulting in insecure passwords [26, 52]. One specific element of PCPs that leads to user frustration while not improving the strength of passwords much is password expiry, which forces users to choose a new password on a regular basis [26, 46, 55]. The removal of this requirement has been discussed by academic research for at least 13 years now [28], but has not been fully implemented by the industry [22, 42].

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.
USENIX Symposium on Usable Privacy and Security (SOUPS) 2023.
August 6–8, 2023, Anaheim, CA, USA

National institutions, such as the National Institute of Standards and Technology (NIST) in the United States of America or the Federal Office for Information Security (BSI) in Germany, are possible facilitators in transferring academic findings into industry practice. These institutions offer recommendations concerning authentication and password policies in particular. Regarding password expiry, NIST removed its suggestion to enforce a regular password change in 2016, and the German BSI followed in 2020.

To understand at what speed such a changed recommendation is implemented in the industry, and especially what problems hinder adoption, we conducted three surveys (2020, 2022, 2023) with German companies after the BSI changed its recommendations. Our survey was based on that of Gerlitz et al. [22], who surveyed German companies in 2019, just before the change mentioned above.

We surveyed the authentication system in use, including detailed questions about the password policy, especially password expiry. We recruited our participants via the BSI newsletter, thus, focusing on companies likely interested in IT security topics.

We found that the number of participants whose companies use a regular password expiry decreased in a statistically significant way from 2019 to 2023. But with 45%, the number of participants whose companies use it is still high. Several of those participants whose companies still use password expiry stated that it is used to increase security, because they do not have the capability to implement the suggested alternative mechanisms as recommended by the BSI, or because someone still requested the change. We discuss these reasons and their implications and offer recommendations for future work and national institutions.

Summarized, our key contributions are:

- We document the progression of authentication processes (PCPs and alternative mechanisms) in German companies over the course of 3.5 years.
- We present reasons why companies do not or cannot comply with the recommendations regarding password

expiry and alternative checks for account compromise.

- We offer suggestions and recommendations for researchers and national institutions.

2 Related Works

In this section, we summarize the published knowledge about the current state of authentication on websites and in companies. We present research that focused on the basis for deciding on these authentication systems and PCPs, and then give a short overview of current academic advice on how PCPs should look like and summarize the current guidelines of NIST and the BSI. We conclude by presenting areas in IT security for which the best practice has changed at some point in time to understand the pace at which such changes can be expected to be implemented.

2.1 Status of Authentication and Basis for Decision

In this section, we look at the current state of authentication in companies and websites and present a paper that looked into the decision process on PCPs.

In 2019, Gerlitz et al. [22] surveyed 83 participants responsible for the authentication process in companies at the time of the study. The authors sought details of their authentication system, such as the type of methods employees can use to log in to their accounts and the PCP in use. The authors found that all companies allowed or demanded passwords for authentication. Most companies specified a minimum length, required specific character classes, and used password expiry for employee passwords. However, when looking at the details of these three components, the actual implementation varied. The most common policy was used by seven companies and required at least eight characters for the password, at least three character classes, and a password expiry of 90 days. Only two participants explicitly mentioned not forcing their employees to change their passwords regularly. Back then, the BSI still recommended password expiry. We build upon this work by replicating the study up to three years later and comparing the results.

In 2022, Hypr, a company aiming for passwordless authentication, conducted interviews with 500 IT decision-makers within financial services organizations in Europe and the United States of America. They found that 32% use 2FA for their employees, while 22% use only the username and password [27].

Research has also looked into password policies on websites and compared them to recommendations by scientists and official institutions. At the beginning of 2019, Gautam et al. [20] extracted and analyzed the password composition policies of 270 websites with the highest ranks in ten different countries. They compared the policies to the recommendations of NIST and found that only around 40% followed the recommended minimum length of 8 or more characters, while more than 70% had an unnecessary maximum length requirement. Two-thirds did not enforce certain character classes and thus followed the recommendations.

Lee et al. [33] analyzed 120 of the most popular websites in 2021 and compared the password policies to current best practices from academia: blocking common passwords, requiring specific character classes, and using a password meter to provide feedback on the security of a password. They found that 60% of the websites do not prevent the usage of the most common passwords at all, and a further 15% seem to use a very limited blocklist, thus still allowing many easy-to-guess and leaked passwords. Contrary to recommendations, 45% of the websites required specific character classes. Only 19% used a password meter of any sort. Interestingly, the authors capture 73 distinct password policies among the 120 websites.

Another paper closely related to our study is that of Sahin et al. [42]. They interviewed eleven website administrators to understand what considerations impact the PCPs they employ and what challenges they face when doing so. The authors found that administrators often face design challenges, competing interests, and deployment challenges. We build upon their work by a) recruiting a larger sample and b) focusing on security professionals within a company. In a professional context, accounts not only hold personal information about a user but might also give access to sensitive company internals. We believe the behavior of decision makers might be influenced by the fact that not only the company's reputation is at stake but also company secrets. This way, it is possible to compare issues in different sectors.

2.2 Current Advice on Password Components

Over the last few years, several researchers have experimented with different elements often seen in password policies, trying to understand their implications on usability and security [25, 30, 45, 49]. Currently, the best combination seems to be one of a minimum length requirement combined with a minimum strength requirement (policies that require the password to exceed a strength threshold) or, if the latter is not possible, a carefully configured blocklist [49].

For a long time, it was recommended that users change their passwords regularly for two reasons: first, if the passwords were ever leaked, chances were high that the list was already outdated when an attacker got access to it. Second, attackers would need more time or computing power to brute force passwords before they were changed. For the latter, 90 days were often seen as a reasonable trade-off that would make it impossible for attackers to guess the passwords in time but still be usable enough for users [48].

The usability part has since then been studied several times, finding that users have trouble recalling new passwords and

find forced password updates annoying [9, 26, 28, 46].

Apart from usability, studies also show that password expiry does not seem to offer the security benefit it was thought to have, and many people simply modify the accounts' current password or reuse a password from another account [8, 26, 46, 55].

While several official recommendations included a regular password change in the past, this has changed now. Instead, other mechanisms should be implemented that check for a compromise, as summarized in the next section.

2.3 Institutional Recommendations Regarding Password Composition Policies

In Germany, where our survey was sent out, the BSI "is the National Security Authority and the chief architect of secure digitalisation in Germany" [17]. Once a year, it publishes an updated version of its IT-Grundschutz Compendium [14], which "offers a systematic approach to information security that is compatible with ISO/IEC 27001" [14]. The BSI additionally hands out implementation hints for some subsections, e.g., for identity and access management, including authentication [16]. Regarding PCPs, the guidelines in the compendium are quite vague and state that "Passwords MUST be sufficiently complex so that they are difficult to guess. Passwords MUST NOT be so complex that users cannot utilise them regularly with a reasonable amount of effort. [15][1]" The implementation hints are more specific and recommend not using words from the personal or work environment and comparing the passwords to lists of leaked passwords. It also suggests combining complexity and length requirements (e.g., 8-12 characters + 4 character classes or 20-25 characters + two character classes).

While in 2019, the compendium included a passage stating "The passwords SHOULD be changed at appropriate intervals" [13], this has changed at the beginning of 2020. Since then, users should only be prompted "to change their passwords with a valid reason. Changes based on the passage of time alone SHOULD be avoided. [15]" Instead, mechanisms must be implemented to detect compromised passwords, e.g., detecting parallel logins from different systems or locations. Only in case these alternative mechanisms cannot be implemented should a regular change be considered [16].

In 2020, the BSI also changed their wording concerning complexity requirements (from "[the organization] MUST specify that only passwords of sufficient length and complexity are to be used [13]" to "Passwords MUST be sufficiently complex so that they are difficult to guess. [15]"), and added a paragraph about two-factor authentication ("[the organization] MUST consider whether passwords are to be used as the sole authentication method, or whether other authentication features or methods may be used in addition to or instead of passwords [15]").

Companies do not have any legal obligation to apply the IT-Grundschutz. Yet, it is one way to implement ISO 27001 in order to get certified [14]. There are several reasons for companies to do this, e.g., reputation, risk calculation, and compliance. Certification is valid for three years, while there are yearly audits [18]. Therefore, changes to the compendium should be implemented in the industry at the latest after three years.

The US American equivalent of the BSI, NIST, recommends that user-chosen passwords should at least be eight characters long and shall be compared "against a list that contains values known to be commonly used, expected, or compromised" [24], e.g., passwords from previous leaks, dictionary words, repetitive or sequential characters, and context-specific words. Since 2016, the recommendations have advised against requiring periodical changes [23]. Instead, verifiers "SHALL implement controls to protect against online guessing attacks" [24], such as risked-based authentication using IP addresses, geolocation, and browser metadata.

2.4 Speed of Adopting Recommendations

After a recommendation is released by, e.g., institutions like NIST, or research, adoption takes time [3, 7, 11, 36] as the people in charge of implementing it are unaware of the recommendation [36], have no time [34, 50], or do not have the necessary knowhow [36]. In many IT security-related topics, technology has changed over the years, and with it, the knowledge of what is secure or usable secure.

To understand how fast the knowledge about best practices takes to reach those in charge of using this information, we highlight this process for two examples: deprecated hashing algorithms and HTTPS.

2.4.1 Deprecated Hashing Algorithms

One security-related area that encounters constant changes in recommendations is hashing. Algorithms get deprecated because they were found to be broken [47] or key lengths have to be increased due to the rising computing power [37]. In 2008, it became clear that MD5 was broken [47] and should thus not be used for storing passwords. Despite this being public knowledge for more than ten years now, some developers are still unaware of this fact. In a 2019 study, Naiakshina et al. [36] asked freelance developers to implement the password storage functionality for a website. Over 20% of the participants used MD5 (18% even stored the passwords in plain text/Base64 encoded). Danilova et al. [10] asked participants to review the program code and included insecure password storage (such as MD5). Only 36% pointed out this problem, and one even mentioned MD5 as a hash algorithm to improve

[1] At the time of writing, the English version of the 2023 version has not been published. The quotes are taken from the English translation of the compendium from 2021. The quoted passages of both German versions are identical.

security. Ntantogian et al. [38] investigated the default password hashing scheme of 25 content management systems and web application frameworks in 2018. MD5 was the default hashing scheme for around 27% of the analyzed CMS. This problem can also be seen when looking at the database of "Have I been pwned" [3]: around 30% of the datasets that included passwords and were breached in 2021 or 2022 used MD5 for password hashing; in 2020, it was 20%.

2.4.2 HTTPS

In 2015, the W3C Technical Architecture Group encouraged the use of HTTPS instead of HTTP [53]. At this point, only around 32% of all pages visited by Firefox browsers supported HTTPS. Around that time, Let's Encrypt was founded and, for the first time, enabled server owners to acquire TLS certificates for free [44]. The updated recommendations, in combination with the easier access to certificates, led to the fact that, in 2022, seven years later, the percentage of servers that supported HTTPS grew to nearly 80% [11], and most (79,8%) of the web servers specifically allow only HTTPS-connections [54].

Even though 80% is the majority, this, in turn, also means that one-fifth of the servers still do not support TLS-secured connections. This can be due to compatibility reasons or an administrator's incorrect mental model of the technology [7, 31].

3 Methodology

We conducted three online surveys recruiting through the BSI mailing list, one in October and November 2020, the second in February and March 2022, and the third in January 2023. The questionnaire and data for 2019 were provided by Gerlitz et al. [22]. For easier readability, we will refer to the surveys and datasets as follows: PCP19 for data presented by Gerlitz et al. [22] that was conducted in 2019, PCP20 for data collected at the end of 2020, and PCP22 and PCP23 for data collected at the beginning of 2022 and 2023.

3.1 Research Questions

The following research questions and hypotheses guide our analysis:

- **RQ1**: How did the authentication system within companies change over the years?
- **RQ2**: How did the usage of a maximum age develop over time after the BSI changed its recommendations in 2020? For this, we had the following hypotheses:
  - ***H1***. *The total number of companies using password expiry decreased from 2019 to 2023.*
    This hypothesis is built on the fact that the BSI dropped their recommendation for using a regular password expiry. Only in case alternative mechanisms, such as checking for parallel logins from different systems or locations, cannot be implemented should a regular forced change be considered. We performed one Fisher's exact test, including all participants who made a statement about their password expiry.
  - ***H2***. *For companies that use password expiry: The time range after which a password is required to be changed increased between 2019 and 2023.*
    We assumed that not all companies could implement alternative checks to remove password expiry entirely. We hypothesized that even if a company cannot remove the password expiry requirement, it would adapt to the BSI recommendation to increase the time intervals between enforced changes. We performed one Wilcoxon rank-sum test and included all participants who used a password expiry and also mentioned a specific time range.
  - ***H3-6***: *Company characteristics or the use of certain policy elements influence whether the companies use a password expiry in 2023.* Factors like time, money, or flexibility could influence adopting the changed recommendations in a company. We, therefore, performed four Fisher's exact tests based on different company characteristics like their size (H3) or whether it belongs to critical infrastructure[2] (H4). We also tested if the usage of checks for password compromises (H5) or if the last change of password policies was before or after 2020 (H6) influenced password expiry usage.
- **RQ3**: What reasons do participants have to still use password expiry?
- **RQ4**: How do companies check for compromised accounts, or what hinders them from implementing such checks?

3.2 Survey Design

The survey consisted of five blocks, as presented in the following paragraphs. Each questionnaire differed slightly from the one before, adapting to new situations and gathering further insights. The survey is given in Appendix A, including annotations highlighting differences between the years.

[2] "Critical infrastructures are organizations or facilities with important significance for the state community, the failure or impairment of which would result in lasting supply bottlenecks, significant disruptions to public safety or other dramatic consequences." Translated from the Federal Office of Civil Protection and Disaster Assistance [39].

Account per Employee and Login The participants were asked whether their company makes use of a centrally-managed account for each employee (Q1[3]), what services it can be used for (Q3a), what authentication methods can be used (Q3b, Q3c), and if two-factor authentication is possible (Q4).

Passwords The password part was shown to all participants who indicated the employee account is secured with passwords. If a company did not have such an account, all questions concerned email passwords. We asked the participants for their password policy (and encouraged them to provide a copy of it, Q6) and for elements allowed or forbidden in the password (Q16, Q17). The participants could indicate how the policies impact the security and usability of the overall authentication system and how often problems occur (Q26 - Q28). Extending the original questionnaire of Gerlitz et al., we also asked the participants whether there are additional specifications for particular user groups (Q18) and when and why the policy was changed last (Q22-Q25). Additionally, the changes in the BSI recommendations were brought up, and the participants were asked whether they looked into the changes and adapted their policy based on them (Q29, Q30). In 2023, we added further open-ended questions to understand if and why a company uses password expiry and how accounts are checked against a compromise (Q8 - Q14).

Biometric Authentication and Hardware Token All participants who indicated that the employee account could be unlocked using biometric authentication or a hardware token were asked for details (which biometric authentication is used (Q34) and whether the token supports FIDO2[4] (Q39)). Similar to the password part, they were asked for their perceived influence of the biometric authentication and token on the security and usability of the authentication system and the frequency of problems (Q35-Q37, Q40-Q42).

Passwordless We added one open-ended question in PCP23 asking for the general sentiments towards passwordless authentication (Q44).

Demographics In the final part, the participants were asked for details about themselves and the company they work for. In PCP19, this included the total number of employees working for the company (Q52) and the number of employees working on IT security topics full-time (Q54). From PCP20 on, the participants were also asked for the sector (Q47), the country the headquarters of the company is located (Q51), whether it can be seen as Critical Infrastructure (Q48), as well as the position of the participant (Q55), and their years of experience (Q56). In all years, the participants indicated their satisfaction with the overall authentication system (Q59). From PCP20 on, they could also suggest that they participated in the survey the previous year (Q58), which only two people reported in PCP22.

[3]The notation Qx references to the corresponding question in our questionnaire.

[4]"Authentication standards based on public key cryptography for authentication" [1]

3.3 Ethics

Our university's Research Ethics Board approved the study, and we adhered to the German data protection laws and the GDPR in the EU. Participants had to consent to their data being used for research before the study began, and we included the option "I don't want to state" for all questions. Participants could drop out at any point in time. The study included multiple open-ended questions. If a participant's answer contained deanonymizing information (e.g., their company name), we deleted it before we continued the analysis.

3.4 Recruitment and Demographics

Participants were recruited through the official newsletter sent by the BSI. Everyone responsible for authentication in a company was invited, and participation was voluntary and not compensated. In this, we followed the original approach by Gerlitz et al. [22], who recruited participants in the same way between September and October 2019. In the latest run in 2023, which also included the highest number of questions, participants took a median time of 16 minutes to finish the survey.

Table 4 and Table 5 in Appendix B show the participants' demographics and their company characteristics.

3.5 Data Quality

We eliminated all incomplete answers from our analysis, as well as one participant whose answers to open-ended questions indicated that they did not understand the questions correctly in PCP19. We further removed the response of one participant whose self-reported role does not clearly include being able to work on the company's authentication in the dataset of PCP20.

Gerlitz et al. [22] not only recruited over the newsletter but also used additional channels to distribute their survey. When comparing the results, we included only those answers by participants recruited through the newsletter for internal validity. To keep the samples as similar as possible, we included only those companies using employee accounts in our analysis. After this filtering, the datasets consisted of 54 (PCP19), 52 (PCP20), 63 (PCP22), and 80 (PCP23) answers. Due to the slightly different filtering, we included fewer participants in our comparison than reported by Gerlitz et al. [22], who reported the data of 83 participants.

All numbers for this filtering process are given in Table 3 in Appendix B.

3.6 Data Analysis

Over the years, the questions in the survey slightly changed. However, all questions for which we compared the results between the years were identical.

3.6.1 Coding

One of the open-ended questions was identical in all years and asked for the password composition policy in use (Q6). Answers for these questions from PCP20 and PCP22 were coded by two authors using the code book that Gerlitz et al. created for PCP19. For this step, we included all the complete answers that we received and merged answers from both years, such that during the coding process, the year in which the answer was given was not known to the coders.

First, both authors coded 31 answers to check for a similar understanding of the code book and then coded additional 31 responses to calculate the inter-coder agreement. After that, each coder then coded half of the answers.

To date, most research on password composition policies has focused on minimum length, password expiry, the number of required character classes (complexity), as well as blocklists (see Section 2.2). To gain a complete overview of these components in the PCPs described above, we decided to code the participants' answers that did not mention their minimum length, password expiry, complexity, or a blocklist as 'not mentioned.'

We had 29 codes and used Recal2 [19] to calculate the inter-coder reliability. For all codes, the reliability lies in the range of (0.47, 1) with a weighted mean of 0.98. Table 6 in Appendix B shows the codes, their occurrences, and the code-specific ICR.

Since we noticed that the answers given for Q6 were very straightforward and did not leave much room for interpretation, the other open-ended responses that we analyzed (Q6 of PCP23, Q8/Q12: Reason for using password expiry, and Q14: How do compromise checks happen or why are they not used of PCP23) were coded by one of the researchers. Two authors discussed all codebooks and answers that were ambiguous.

We proceeded the same way for answers given for the "other"-option in multiple-choice questions. All citations from these answers in this paper are translated from German.

3.7 Limitations

This work needs to be interpreted in light of the following limitations: Even though we recruited the participants through the BSI newsletter and clearly stated who the survey is aimed at, we cannot be sure that only the responsible employee took part. From PCP20 on, participants were asked what position they held. Except for one, all participants who specified their role indicated working in a position with detailed domain knowledge about the company's authentication system (see Table 5). Yet especially for bigger companies, we cannot rule out that only one member of the IT security team took part in the survey. We checked for duplicates in the company characteristics in combination with the given PCP but could not identify any identical entries.

All data are based on self-reports, and participants might have forgotten to include elements, especially for the password composition policy. We tried to counter this by asking them to copy and paste their policy.

Using the newsletter sent by the BSI for recruitment may have caused that in our samples, the participants are a) already interested in security topics and news and b) an even more interested subgroup, as those are more likely to read the study invitation and follow it. We discuss the possible implications in Section 5.3.

Over the years, the survey was adapted, and questions were added. This could have caused participants to be in slightly different states of mind when answering questions. However, we took care to include new questions only after similar questions were already asked and included page breaks.

The survey for PCP22 contained minor improvements mentioned above, as well as an additional question about the reason for the existence of a password expiry that was not recommended by the BSI anymore at this point. After the newsletter inviting participants for PCP22 was sent and 36 participants had already completed it, we noticed that the questionnaire for PCP20 was handed out by mistake that did not include the changes and the additional question. We still decided to switch to the improved survey since we were confident that the additional insights from the new questions outweighed the minimal risk of receiving different answers due to the slightly modified questions.

4 Results

In this section, we present the findings of our survey. The changes between the years of the used authentication systems within companies are presented in Section 4.1 (RQ1). The usages of password expiry (RQ2) are summarized in Section 4.2. In Section 4.3 (RQ3), we present the reasons participants gave for using a password expiry, and Section 4.4 (RQ4) summarizes how companies currently check whether accounts are compromised and what issues hinder them in implementing checks.

4.1 RQ1 - Evolution of Authentication Systems

This section considers the evolution of authentication methods and password composition policies between 2019 and 2023.

4.1.1 Possible Authentication Methods

Figure 1 shows the percentage of participants per year who stated that their company offers their employees to authenticate using passwords, biometric authentication, and hardware tokens, independent of their usage as the primary or secondary factor. The use of authentication methods apart from passwords has risen steadily over the last few years. In 2023 for the first time, two companies indicated that their company does not use passwords. All numbers can be found in Table 2 in Appendix B.

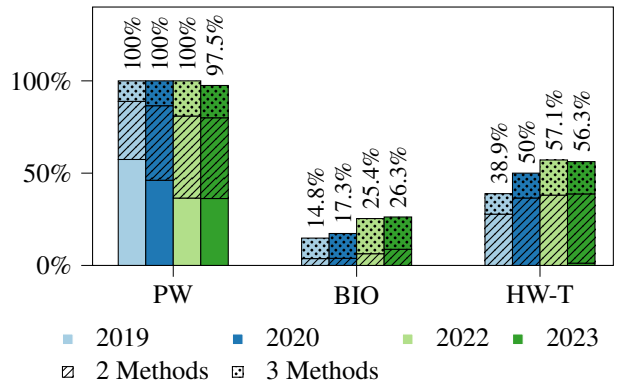


Figure 1: Percentage of participants that indicated that their company enables authentication using passwords (PW), biometric authentication (BIO), or hardware token (HW-T). Using authentication methods other than passwords rose steadily over the years. The bars also indicate the number of companies using one, two, or three authentication methods. Biometrics are seldom used as secondary factor alone.

If participants indicated that more than one authentication method could be used, they were asked whether these methods needed to be used in combination (two-factor authentication). Starting in PCP20, participants who mentioned only one method were additionally asked whether there is any possibility for two-factor authentication. The number of those requesting 2FA also increased, at least between 2019 and 2022: 22.2% of the companies that participated in 2019 required two-factor authentication – 32.7% did so in 2020, 55.6% in 2022, and 51.2% in 2023.

4.1.2 Usage of Password Components

Participants were asked to indicate their current password policy (Q6). This question text included the example, "e.g., at least x characters, new password needs to be selected after x days," and participants were encouraged to provide a copy of their policy. In PCP23, we also explicitly asked for password expiry in a separate question that was shown after a page break after Q6.

Figure 2 shows the development of the complexity, password expiry, and minimum length from PCP19 to PCP23. There is a slight increase in the number of participants who mentioned that their company requires all four character classes to be used in the passwords. While NIST advises against such complexity requirements, the BSI currently includes complexity requirements in their implementation hints (see Section 2.3).

The figures also show a trend towards longer passwords and larger time ranges before the passwords expire, e.g., the number of companies that require a password change every 90 days decreased from 33.3% in 2019 to 6.2% in 2023, while those who explicitly mentioned not using password expiry rose from 1.9% in 2019 to 22.5% in 2023. In 2023, 10.0% of the participants did not include any information about password expiry in the open-ended response but disclosed they actually use a password expiry in the question explicitly asking for it (Q11). We further explore the details of password expiry in Section 4.2. The concrete numbers of companies using a certain minimal length, complexity, and password expiry are shown in Table 6 in Appendix B.

We also looked at the most common combination of password expiry, a required minimum length, and complexity for each year and show the results in Table 1. The most common combination in PCP19 and PCP20 (minimum length of 8 characters, enforcing three character classes, and using a password expiry of 90 days) was not used by any participant in PCP23.

Summary for RQ1 From 2019 to 2023, the companies our participants worked for offered more authentication methods next to passwords, and the number of participants whose companies require 2FA rose by almost 20 percentage points. Looking at PCPs, it is now more common than in 2019 to require longer passwords (12 or even 14 characters) and to refrain from using password expiry.

4.2 RQ2 - Password Expiry

As detailed in Section 2.3, the BSI advises against a forced frequent change of passwords. Instead, companies should run analyses on whether a user account is compromised. This could, for example, be done by checking for parallel logins from several systems or locations. A regular change should only be considered if such checks are impossible.

While the development of the days after which a password change is required is given in the previous section, this section investigates H1 and H2, and we take a closer look at company characteristics that may indicate the use of password expiry (H3-6).

Since we wanted to dive deeper into the analysis of password expiration, we added an additional closed question about password expiry in PCP23 to double-check the open-ended general PCP question.[5] When comparing the numbers from

[5]We added a page break to ensure participants would not be influenced.

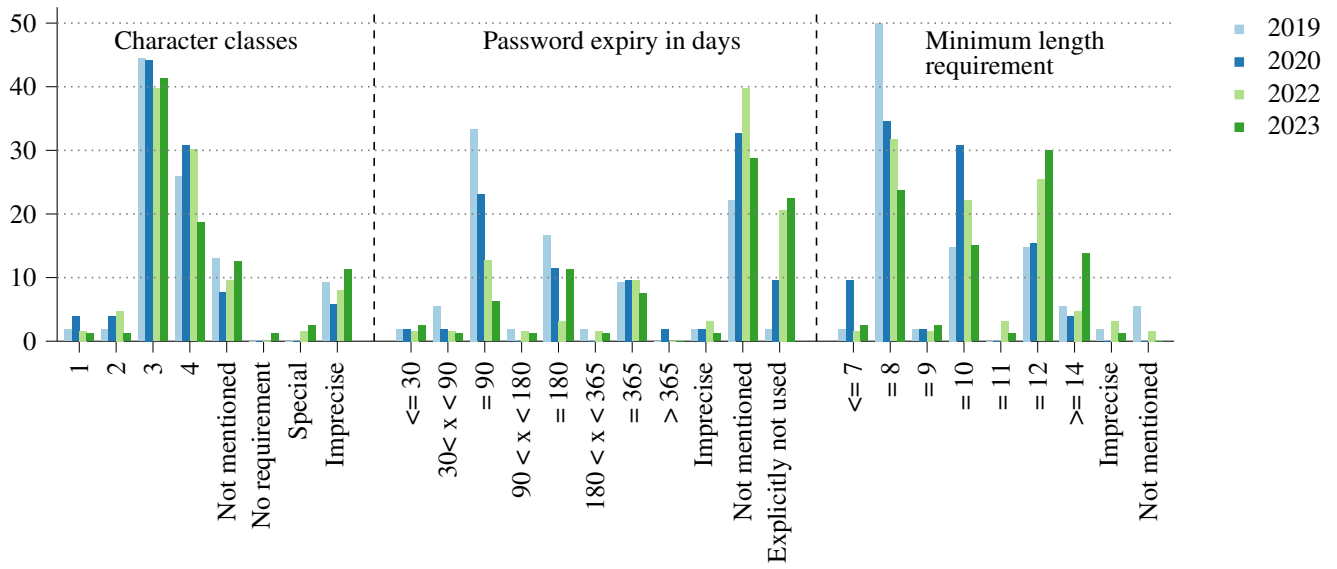


Figure 2: Overview of the required number of character classes that need to be covered in the user's passwords, number of days after which a user is requested to change their password (password expiry), and required minimum length in 2019, 2020, 2022, and 2023. Y-axis shows percentages. Answers from Q6. Findings: **1) Character classes:** Participants whose companies require at least three character classes either mentioned certain classes or allowed a password as long as any three classes were used. There are no big differences between the years. In their implementation hints, the BSI gives examples that include enforcing several character classes (see Section 2.3). **2) Password expiry:** The number of companies using password expiry of 90 or 180 days decreased from 2019 to 2023, and more participants explicitly mentioned not to use password expiry in 2022, following the BSI recommendations. **3) Minimum length:** Requiring passwords with at least 10, 12 or 14 characters is more common in 2023 than it was in 2019, where most PCPs asked for at least eight characters.

the closed question to the open-ended answers, we noticed that ten percentage points fewer participants mentioned password expiry in the open-ended question than in the closed question (35.0% vs. 45%). It is interesting to note that 10% did not think to mention password expiry when describing their policy.

While we cannot say this for sure, we think it is likely that there has been a similar amount of underreporting in the previous years. But to be on the safe side, when we compare PCP23 to previous years, we use only the open-ended data, so the comparisons are made based on the same measurement instrument, i.e., we only use the 35.0% that were captured the same way as in PCP19.

However, we believe 45% to be more accurate and use it wherever applicable.

4.2.1 RQ2 - H1

We hypothesized that the *total number of companies using password expiry decreased from 2019 to 2023.* To investigate this question, we performed two Fisher's exact tests based on answers given to Q6. The tests are based on slightly different assumptions: For the first one, we only included those participants who explicitly said something about their password expiry: either mentioning a time span or indicating they do not use one. In PCP23, 26.9% of those participants who did not mention a password expiry in the open-ended question (Q6) later stated one in the closed question specifically asking for it. The results show a statistically significant difference between the results from 2019 ($n_{exp19} = 39$, $n_{noexp19} = 1$) to 2023 ($n_{exp23} = 28$, $n_{noexp23} = 18$): $p_{(19,23)} = 0.00$, OR = 25.07.

For the second test, we included all participants and assumed that not mentioning expiry is the same as not having one. In PCP23, this was true for 73.1% of those who did not mention password expiry in Q6. This test also shows a statistically significant difference between the results from 2019 ($n_{exp19} = 39$, $n_{noexp19} = 13$) to 2023 ($n_{exp23} = 28$, $n_{noexp23} = 44$): $p_{(19,23)} = 0.00$, OR = 4.71.

This suggests that within three years after the change, there has been a shift towards the new recommendations.

4.2.2 RQ2 - H2

We further hypothesized that *for companies that use password expiry: The time range after which a password is required to be changed increased between 2019 and 2023.* For this analysis, we included only participants who used password expiry. We excluded all participants who gave no clear answer about the length of their password expiry (e.g., only stating that there is a password expiry but giving no numbers.) We

Expiry (days)	Min. Len	Complexity	2019	2020	2022	2023
90	8	3	5	4	3	0
90	8	4	4	3	1	0
90	10	3	3	1	1	0
180	8	3	3	0	0	2
180	8	4	1	3	0	1
Not mentioned	8	3	2	1	4	2
Not mentioned	8	4	3	3	3	2
Not mentioned	12	3	0	3	1	5
Not mentioned	12	4	2	2	6	2
Explicitly not	12	3	0	0	1	4

Table 1: Number of times a certain policy was mentioned in 2019, 2020, 2022, and 2023. The more participants mentioned their company uses a policy, the darker the cell is shaded. A PCP is included in this table if it appeared at least 3 times in at least one year. While the most common policies in 2019 included a password expiry, this trend has shifted in 2023, where the most commonly mentioned policies do either not mention regular password rotation or explicitly state not to use one.

performed a Wilcoxon rank-sum test for which we included 38 answers from 2019 and 27 from 2023. Figure 2 showed a trend towards fewer companies that enforce a password change after 90 or 180 days, and this theme was also picked up in the open-ended responses: "Jumping to unlimited passwords takes time because technical change and culture change take time. We went from 90 days to 1 year." Yet, the tests did not show a statistically significant result ($p_{(19,23)}$ = 0.131, $Z_{(19,23)}$ = -1.511).

4.2.3 RQ2 - H3-6

Finally, we were interested in whether *company characteristics or the use of certain policy elements influence whether the companies use a password expiry in 2023.* We conducted four Fisher's exact tests to analyze the impact of the following variables on the existence of a password expiry: company size (<500, >=500), critical infrastructure (no, yes), last policy change (before BSI changes, after BSI changes), and technical measures that check for account compromise (no, yes). We used data from the open-ended response (Q6) and the explicit question asking for password expiry (Q11) for these tests. Taken together, 45% of the participants mentioned that their company uses a password expiry in 2023.

After correcting the results using a Bonferroni-Holm correction, none of the tested factors remained to have a statistically significant impact on the existence of password expiry. The contingency table for all tests is given in Table 9 in Appendix B.

Summary for RQ2 Over 40% of the participants stated that their company still uses password expiry in 2023, although this is statistically significantly less than in 2019. None of the characteristics for which we tested differences in the use of password expiry showed a statistically significant result.

4.3 RQ3: Why do Companies Require a Regular Password Change?

Thirty-six participants in PCP23 and nine in PCP22 answered why their company uses password expiry.

Apart from this explicit question, some participants included a reason for using password expiry in their open response to Q6.

In the following, we present the most commonly mentioned themes. The coding table is given in Table 7 in Appendix B.

Increase Security In 2023, 17 participants said their company uses a password expiry for security reasons. Half of them stayed vague and stated "[password expiry] gives *some* security" or "improvement of password security." The remaining eight mentioned more specific reasons, e.g., "Because after a year, the risk of misuse of the PW through reuse with other, potentially corrupted services, is too great."

In 2022, two participants stated their company uses a password expiry to get rid of lost ($n = 1$) or leaked passwords where employees did not follow the recommended change after being notified about the leak ($n = 1$).

Still Demanded Another commonly mentioned reason was that someone or some regulation demands a password expiry. Four participants stated their CEO or internal policies require this regular change, "[...] contrary to the recommendation of the IT sec [department]" or because of "inertia in our security standards." Three participants mentioned that customers or the customers' compliance requires regular changes, and five participants mentioned official requirements, e.g., "Official (in my eyes outdated) requirements in handling officially classified data." The latter two reasons were also mentioned in previous years.

No Alternatives A small number of participants (Four in 2023, one in 2022) use password expiry because they have "No other method implemented yet" or because "there is currently no other technical solution." We look deeper into these alternatives in Section 4.4.

Best Practice Some participants (Two in 2023 and two in 2022) stated that password expiry is best practice or recommended by the BSI. One participant stated: "[We] consider it wrong not to change [the password] regularly." This was already a theme in 2022, so we asked in 2023 whether participants perceived a regular change as best practice before asking why password expiry was used. This question was affirmed by 23.1% of the PCP23 participants.

Currently Changing We also saw participants (One in 2023 and one in 2022) who stated they were currently in the

process of eliminating password expiry. One mentioned: "The auditor still has to be convinced."

4.4 RQ4: How do Companies Check for Compromised Accounts and What Hinders Them?

The BSI specifies that mechanisms must be implemented to detect compromised passwords (see Section 2.3). So, we were interested in the mechanisms companies use or whether they have problems implementing them. We included this question in the PCP23 survey, and 66 participants answered it. The coding table is shown in Table 8 in Appendix B.

Of those 66 answers, 36 participants said their companies use technical solutions to check for password or account compromises, whereas 25 do not. In three cases, the participants gave no clear answer, and in two cases, checks are not used consistently for all systems (e.g., used for SSO, but not for AD).

Those who use checks most often do so by checking against databases of leaked passwords ($n = 8$) or by using tools that check for anomalies within the system or during logins ($n = 16$).

The absence of technical checks was most often explained by missing resources (time, financial, or human resources) ($n = 7$), by structural and organizational issues ($n = 5$), or because the technical solution was not straightforward ($n = 9$). One participant mentioned a "conflict with the works council." Specific problems mentioned were that third-party tools are needed or behavior-based detection tests lead to several false positives in the past.

Three participants questioned the possibility of checking for compromises in general: "It is not clear to us how to check if a [password] is 100% not compromised."

Additional ($n = 2$) or alternatively ($n = 4$) to technical checks, some participants stated their companies encourage their employees to check for a compromise themselves.

Summary for RQ3 and RQ4 Companies still use password expiry in 2023, mainly to increase IT security or because another entity required it. Technical measures that check for account compromise were often not implemented because of missing resources.

5 Discussion

We conducted three surveys over three years to understand how password composition policies evolve. We found that companies now require passwords to have more characters than in 2019 and found significantly fewer participants whose companies rely on password expiry. When specifically asking for the reason for still using expiry in PCP23, we found IT security to be one of the leading explanations.

This section discusses the reasons why companies still use password expiry, draws connections to related work, and gives recommendations for future work and policymakers.

5.1 Password Expiry and IT Security

The BSI started advising against password expiry at the beginning of 2020. With this, they finally followed scientific work, a full decade after it showed the usability of password expiry to be a problem [9, 28, 46]. In our study, four participants explicitly mentioned that employees wrote passwords down when they had to change them regularly, and the company thus got rid of this requirement.

Nonetheless, in the latest survey run in 2023, still, 45% of the participants stated that their company forces regular password changes from their employees. Several of them (17) argued for an improvement in IT security, confirming the findings of Sahin et al. [42]. In the following, we discuss these reasons and evaluate whether this situation is problematic.

Compensation of Technical Issues We found cases where password expiry was used to compensate for other technical issues. One participant mentioned using expiry to get rid of old hashing algorithms, one referred to a maximum password length of eight that was set by the system, and a third explained that MFA was not yet implemented for all accounts. It can make sense to keep password expiry to bridge the time until new technologies are implemented (as mentioned in the latter example). However, accepting all the drawbacks that regular password changes bring because of a legacy system that itself can be a security risk seems like a missed opportunity: Instead of adopting to constraints of a system, system administrators could argue that they need a new and more secure system that can fulfill the updated recommendations. However, we must acknowledge that business constraints can make this difficult.

Initiate Password Development The most commonly used arguments for password expiry concerned initiating the development of passwords, i.e., making sure that a password is not in use anymore when an attacker gets access to it and reducing the problems that come with password reuse by, e.g., decoupling the employee account from private accounts for which the employee used the same password.

Both of those arguments sound reasonable in theory, especially as credential stuffing attacks (where an attacker tries to log into accounts by using passwords associated with that user in leaks) made up almost a third of all attacks across the Arkose Labs network [32] and more than 12 billion leaked login combinations are public by the time of writing in 2023 [3].

However, estimating the real security benefit of such changes is hard. Studies indicated that many individuals simply modify a previous password when being forced to change it [26, 55]. This makes it easy for an attacker to guess the new password if they have access to a previous one, especially

when considering advanced attacks where not only the password that is included in leaks is used but also variations of it (e.g., following the approach by Pal et al. [40]). Yet, simple attacks that use exactly the same password as found in leaks might be prevented.

Further, many attacks, e.g., data theft, do not require much time, and according to a study by Agari, 50% of the credentials were used within twelve hours after they were compromised [6]; thus, requiring password changes every month will likely not prevent many attacks.

If persistence is the goal, attackers will attempt to create further footholds so that losing access to the first account does not lock them out.

And while the results from a survey study by Habib et al. [26] indicate that users do not seem to choose weaker passwords when updating them compared to creating new ones, Adams and Sasse [5] argue that a regular password change could lead users to create very simple passwords. On a larger scale, it is unclear whether users' strategies for initially creating the password differ, depending on whether they know they will have to change it soon or can keep it for a longer period.

What is clear is that physical password handling differs depending on whether passwords need to be changed frequently or not: Habib et al. [26] saw a statistically significant increase in storing the main workplace password in the web browser when password expiry was in use. Depending on the details of this (usage of a browser password manager, behavior concerning device locking when leaving the desks, etc.), this might have a positive or negative impact on security. Similarly, if users are also more likely to write down their passwords on sticky notes, physical access to a workspace would be a problem. Yet, unobserved access to a workspace might come with several other risks anyway, such as being able to insert a physical key logger (even though this involves more planning than simply using the password from a sticky note).

Adding "Some" Security The uncertainty of how much security is actually added was also present in the answers, where participants associated password expiry with "*some* security". While it can be reasonable to use every opportunity, even the small ones, to increase IT security, people nowadays already have to deal with many security mechanisms in their work environment (e.g., authentication, secure messaging, physical access control). Removing those requirements that are very time-consuming [9], and can even be replaced by technical checks, thus seems like a good idea.

5.2 Further Reasons for Delayed PCP Updates

In this section, we discuss arguments for using password expiry apart from increasing IT security.

Alternative Mechanisms Cannot be Implemented One participant in 2022 mentioned that their company is not able to remove the requirement for a regular change because of "inadequate control mechanisms to detect compromise." In 2023, participants mentioned technical reasons, e.g., that third-party tools are needed, current tools lead to too many false positives, or that, in general, the implementation of such checks is "too complex."

This problem can also be seen in other security-related areas, such as deploying updates, where systems are not updated because of legacy software that is known to be incompatible with up-to-date environments [34, 50].

The area of these alternative mechanisms remains to be studied in more detail. Further reasons for some companies not being able to use checks that indicate a password compromise need to be identified. One paper that has already studied these alternative mechanisms was published by Markert et al. [35]. The authors studied administrators' understanding of risked-based authentication. They found that their participants struggled with the meaning of the given risk levels and the configuration interface in general.

Depending on further findings, it might be helpful if the BSI, or any institution with a similar influence, could publish additional information about the available alternatives and how they can be implemented. At the point of writing, the suggestions concerning alternative mechanisms in the current implementation hints are very vague: "For example, logging and the corresponding evaluation of log files can be used to determine whether there have been unusual accesses or hacking attempts. Special security products are also available for databases, operating systems, web servers, and other applications."[6]

Looking at the area of HTTPS, the increase in its usage was not only sparked by a changed recommendation but also because there was a new and very easy way to follow it (see 2.4).

Inertia Participants pointed to the necessity to follow requirements that still demand regular change, e.g., federal offices or the PCI (Payment Card Industry Data Security Standard). In some cases, internal security standards require the use. As pointed out in Section 2.4 and also mentioned by the participants, changes take time to reach every stakeholder, and one participant mentioned the word *inertia*.

The problem of contradictory guidelines can appear whenever multiple institutions publish different recommendations for the same topic. In these cases, administrators must decide which guidelines to follow or how to combine them.

No Knowledge About Change and Misconceptions Many technical news portals, e.g., [41, 43] and newspapers, e.g., [12, 29] reported the removal of password expiry in the new BSI

[6] Translated from the German implementation hints [16]

recommendations. However, we still encountered one participant who stated that this is recommended by the BSI. In 2020, 25% of the participants mentioned being unaware of the BSI changes published eight months earlier while being subscribers to the newsletter. This phenomenon can also be seen in other areas (see section 2.4).

We noticed misconceptions about how checks for account compromise happen and potentially problematic views toward security in general. One participant, e.g., stated that passwords need to be shared with third-party tools for compromise checks. Although this might be the case for some checks, there are solutions that circumvent this problem [2].

Another participant mentioned that it is never 100% sure whether a password is compromised. While this is true for almost any other security-related topic, it should not lead to not using compromise checks at all.

Here, efforts such as the Let's Hash website of Geierhaas et al. [21] can help by providing participants with a programming aid that supports developers in securely implementing password storage. It seems beneficial to further go this route and offer code-fulfilling recommendations such as from the BSI, NIST, or OWASP. While such a website is a good starting point, the people implementing the recommendations still need to be made aware of such platforms and that their knowledge is not up to date.

Processes Need to be Observed Depending on the size and structure of a company, changing the PCP can require potentially complex processes and involve multiple parties. In our sample, we found cases where the CEO was involved in creating the PCP. While we do not know the whole picture, we sometimes assume a clash of very different goals. A CEO of a small company stated: "From our point of view, the management is responsible for the guidelines and not the IT admins." In another case, the CEO of a company required password expiry, even though the security department advised against this. In cases like this, non-technical explanations of why something should or should not be used from the IT perspective might help to convince decision-makers without deep technical background and help to mediate between different stakeholders.

Apart from this decision process, a new policy has to be included in the systems. Several participants indicated they use Microsoft's Active Directory for authentication. It would be interesting to ascertain whether and how the usage of central software positively affects processes in general. This way, and in combination with secure defaults, the processes might be improved in simplicity, speed, and security, as has already been seen in other areas [4, 51].

Arguments for Change and Prioritization One of the most commonly mentioned reasons why checks for password compromise are not implemented are missing resources, i.e., time, money, or human resources.

If decisions need to be made for prioritizing tasks, low-priority tasks are often postponed, and other stakeholders must be convinced that allocating time for this is a good idea. For this reason, the arguments for a change that impacts usability more than security need to be good. Official recommendations could explain their rationale behind decisions, perhaps even beyond the security topic. That way, decision-makers have a better overview, and it might be easier to understand and communicate possible implications.

5.3 Sample and Recruitment Bias

We recruited over the newsletter sent by the BSI, thus focusing on companies interested in security-related topics, either out of an employee's personal interest or because their company has to follow specific guidelines. The latter might be the case for the 13.8% in our sample (PCP23) that indicated their company can be seen as critical infrastructure and for those 27.5% who indicated their company is certified with a certification relevant for IT security (6.2% indicated both). Yet, the effect of such a security focus is unclear and could lead to two very distinct outcomes: either following recommendations very closely or using every opportunity to increase security, even if the measures taken are questionable in terms of improving security.

6 Conclusion

In 2020, the BSI changed its guidelines regarding password composition policies and removed the advice to include password expiry. Instead, they recommend only enforcing a change if a password is compromised.

We conducted three survey studies after the guideline change to understand how fast and well these suggestions are implemented and what problems might hinder an adoption.

We found that fewer companies require a regular password change ofter the years (72.2% in 2019 to 45% in 2023). Participants who still use a regular expiry explained this with security improvements and additional guidelines by other institutions they must follow. Alternative checks were often not implemented because of missing resources or because of technical hurdles. We believe that it might be helpful if decision-makers were supported with more precise information about these alternatives than what is currently included in the guidelines given by the BSI.

Acknowledgments

We thank the Werner Siemens-Stiftung (WSS) for their generous support of this project. We thank Julia Angelika Grohs, Bilal Kizilkaya, Charlotte Theresa Mädler, and our anonymous reviewers for their help and feedback.

References

[1] FIDO2 - FIDO Alliance. https://fidoalliance.org/fido2/, No year given. Accessed: May 26, 2023.

[2] Have I Been Pwned: API v3. https://haveibeenpwned.com/API/v3, No year given. Accessed: May 26, 2023.

[3] Have I Been Pwned: Pwned websites. https://haveibeenpwned.com/, No year given. Accessed: May 26, 2023.

[4] Sigstore. https://www.sigstore.dev/, No year given. Accessed: May 26, 2023.

[5] Anne Adams and Martina Angela Sasse. Users Are Not the Enemy. *Commun. ACM*, 42(12):40–46, December 1999.

[6] Agari. Anatomy of a compromised account. https://www.agari.com/resources/guides/anatomy-compromised-email-account, No year given. Accessed: May 26, 2023.

[7] Devdatta Akhawe, Johanna Amann, Matthias Vallentin, and Robin Sommer. Here's My Cert, so Trust Me, Maybe? Understanding TLS Errors on the Web. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13, page 59–70, New York, NY, USA, 2013. Association for Computing Machinery.

[8] Sonia Chiasson and Paul C Van Oorschot. Quantifying the security advantage of password expiration policies. *Designs, Codes and Cryptography*, 77:401–408, 2015.

[9] Yee-Yin Choong, Mary Theofanos, and Hung-Kung Liu. United States Federal Employees' Password Management Behaviors - a Department of Commerce case study. Technical Report NIST IR 7991, National Institute of Standards and Technology, April 2014.

[10] Anastasia Danilova, Alena Naiakshina, Anna Rasgauski, and Matthew Smith. Code Reviewing as Methodology for Online Security Studies with Developers - A Case Study with Freelancers on Password Storage. In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*, pages 397–416. USENIX Association, August 2021.

[11] Let's Encrypt. Let's Encrypt - Stats. https://letsencrypt.org/stats/, 2022. Accessed: May 26, 2023.

[12] FAZ. Nicht immer wieder das Passwort ändern. https://www.faz.net/aktuell/wirtschaft/bsi-verabschiedet-sich-vom-regelmaessigen-passwort-wechsel-16616730.html, 2020. Accessed: May 26, 2023.

[13] Federal Office for Information Security (BSI). IT-Grundschutz Compendium - Final Draft, 1 February 2019. https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/Grundschutz/International/bsi-it-gs-comp-2019.pdf?__blob=publicationFile&v=1, 2019. Accessed: May 26, 2023.

[14] Federal Office for Information Security (BSI). BSI - IT-Grundschutz. https://www.bsi.bund.de/EN/Themen/Unternehmen-und-Organisationen/Standards-und-Zertifizierung/IT-Grundschutz/it-grundschutz_node.html, 2021. Accessed: May 26, 2023.

[15] Federal Office for Information Security (BSI). IT-Grundschutz Compendium - Final Draft, 1 February 2021. https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/Grundschutz/International/bsi_it_gs_comp_2021.pdf?__blob=publicationFile&v=4, 2021. Accessed: May 26, 2023.

[16] Federal Office for Information Security (BSI). Umsetzungshinweise zum Baustein: ORP.4. Identitäts- und Berechtigungsmanagement. https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/Grundschutz/Umsetzungshinweise/Umsetzungshinweise_2021/Umsetzungshinweis_zum_Baustein_ORP_4_Identitaets_und_Berechtigungsmanagement.pdf?__blob=publicationFile&v=1, 2021. Accessed: May 26, 2023.

[17] Federal Office for Information Security (BSI). BSI - Organisation and structure. https://www.bsi.bund.de/EN/Das-BSI/Organisation-und-Aufbau/organisation-und-aufbau_node.html, No year given. Accessed: May 26, 2023.

[18] Federal Office for Information Security. Zertifizierung nach ISO 27001 auf der Basis von IT-Grundschutz. https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/Grundschutz/Zertifikat/ISO27001/Zertifizierungsschema_Kompendium.pdf?__blob=publicationFile&v=1, 2019. Accessed: May 26, 2023.

[19] Deen Freelon. ReCal2. http://dfreelon.org/utils/recalfront/recal2/, No year given. Accessed: May 26, 2023.

[20] Anuj Gautam, Shan Lalani, and Scott Ruoti. Improving Password Generation Through the Design of a Password Composition Policy Description Language. In *Eighteenth Symposium on Usable Privacy and Security (SOUPS 2022)*, pages 541–560, Boston, MA, August 2022. USENIX Association.

[21] Lisa Geierhaas, Anna-Marie Ortloff, Matthew Smith, and Alena Naiakshina. Let's hash: Helping developers with password security. In *Eighteenth Symposium on Usable Privacy and Security (SOUPS 2022)*, pages 503–522, Boston, MA, August 2022. USENIX Association.

[22] Eva Gerlitz, Maximilian Häring, and Matthew Smith. Please do not use !?_ or your license plate number: Analyzing password policies in german companies. In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*, pages 17–36. USENIX Association, August 2021.

[23] Paul A. Grassi. Publish#1. https://github.com/usnistgov/800-63-3/commit/3fb942e2f795f681155144d06933db28f29278e0#diff-c10a8efb34bad3a0b9a47880106e0f255f5f866f12fdcccbd73b7338ea82d301, 2016. Accessed: May 26, 2023.

[24] Paul A Grassi, James L Fenton, Elaine M Newton, Ray A Perlner, Andrew R Regenscheid, William E Burr, Justin P Richer, Naomi B Lefkovitz, Jamie M Danker, Yee-Yin Choong, Kristen K Greene, and Mary F Theofanos. Digital identity guidelines: authentication and lifecycle management. Technical Report NIST SP 800-63b, National Institute of Standards and Technology, Gaithersburg, MD, June 2017.

[25] Hana Habib, Jessica Colnago, William Melicher, Blase Ur, Sean Segreti, Lujo Bauer, Nicolas Christin, and Lorrie Cranor. Password creation in the presence of blacklists. In *Workshop on Usable Security (USEC) 2017*, page 50, 2017.

[26] Hana Habib, Pardis E. Naeini, Summer Devlin, Maggie Oates, Chelse Swoopes, Lujo Bauer, Nicolas Christin, and Lorrie F. Cranor. User behaviors and attitudes under password expiration policies. In *Fourteenth Symposium on Usable Privacy and Security (SOUPS) 2018*, pages 13–30, Baltimore, MD, August 2018. USENIX Association.

[27] Hypr. Report: State of Authentication in the Finance Industry 2022. https://get.hypr.com/state-of-authentication-in-the-finance-industry-2022, No year given. Accessed: May 26, 2023.

[28] Philip G. Inglesant and Martina A. Sasse. The True Cost of Unusable Password Policies: Password Use in the Wild. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, page 383–392, New York, NY, USA, 2010. Association for Computing Machinery.

[29] Jurik Caspar Iser. Bundesamt hält regelmäßigen Passwortwechsel nicht mehr für notwendig. https://www.zeit.de/digital/datenschutz/2020-02/bsi-empfehlung-passwort-wechsel, 2020. Accessed: May 26, 2023.

[30] Patrick G. Kelley, Saranga Komanduri, Michelle L. Mazurek, Richard Shay, Timothy Vidas, Lujo Bauer, Nicolas Christin, Lorrie F. Cranor, and Julio Lopez. Guess again (and again and again): Measuring password strength by simulating password-cracking algorithms. In *2012 IEEE Symposium on Security and Privacy*, pages 523–537. IEEE, 2012.

[31] Katharina Krombholz, Karoline Busse, Katharina Pfeffer, Matthew Smith, and Emanuel von Zezschwitz. "If HTTPS Were Secure, I Wouldn't Need 2FA" - End User and Administrator Mental Models of HTTPS. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 246–263, 2019.

[32] Arkose Labs. 70% Increase in Fraudulent Account Registrations Puts Digital Accounts in the Crosshairs, Arkose Labs Reports. https://www.businesswire.com/news/home/20210805005657/en/70-Increase-in-Fraudulent-Account-Registrations-Puts-Digital-Accounts-in-the-Crosshairs-Arkose-Labs-Reports, 2021. Accessed: May 26, 2023.

[33] Kevin Lee, Sten Sjöberg, and Arvind Narayanan. Password policies of most top websites fail to follow best practices. In *Eighteenth Symposium on Usable Privacy and Security (SOUPS 2022)*, pages 561–580, Boston, MA, August 2022. USENIX Association.

[34] Frank Li, Lisa Rogers, Arunesh Mathur, Nathan Malkin, and Marshini Chetty. Keepers of the machines: Examining how system administrators manage software updates for multiple machines. In *Fifteenth Symposium on Usable Privacy and Security (SOUPS 2019)*, pages 273–288, Santa Clara, CA, August 2019. USENIX Association.

[35] Philipp Markert, Theodor Schnitzler, Maximilian Golla, and Markus Dürmuth. "As soon as it's a risk, I want to require MFA": How Administrators Configure Risk-based Authentication. In *Eighteenth Symposium on Usable Privacy and Security (SOUPS 2022)*, pages 483–501, Boston, MA, August 2022.

[36] Alena Naiakshina, Anastasia Danilova, Eva Gerlitz, Emanuel von Zezschwitz, and Matthew Smith. "If You Want, I Can Store the Encrypted Password": A Password-Storage Field Study with Freelance Developers. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–12, New York, NY, USA, 2019. Association for Computing Machinery.

[37] NIST Special Publication 800-57 Part 1: Recommendation for Key Management. https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-57pt1r5.pdf, 2020. Accessed: May 26, 2023.

[38] Christoforos Ntantogian, Stefanos Malliaros, and Christos Xenakis. Evaluation of password hashing schemes in open source web platforms. *Computers & Security*, 84:206–224, 2019.

[39] Federal Office of Civil Protection and Disaster Assistance. Kritische Infrastrukturen - BKK. https://www.bbk.bund.de/DE/Themen/Kritische-Infrastrukturen/kritische-infrastrukturen_node.html, No year given. Accessed: May 26, 2023.

[40] Bijeeta Pal, Tal Daniel, Rahul Chatterjee, and Thomas Ristenpart. Beyond Credential Stuffing: Password Similarity Models Using Neural Networks. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 417–434, May 2019.

[41] Dieter Petereit. BSI rät jetzt von regelmäßigem Passwort-Wechsel ab. https://t3n.de/news/bsi-raet-regelmaessigem-ab-1249147/, 2020. Accessed: May 26, 2023.

[42] Sena Sahin, Suood Al Roomi, Tara Poteat, and Frank Li. Investigating the Password Policy Practices of Website Administrators. In *2023 2023 IEEE Symposium on Security and Privacy (SP) (SP)*, pages 1437–1454, 2023.

[43] Jürgen Schmidt. Passwörter: BSI verabschiedet sich vom präventiven, regelmäßigen Passwort-Wechsel. https://heise.de/-4652481, 2020. Accessed: May 26, 2023.

[44] Seth Schoen. Let's Encrypt Brings Free HTTPS to the World: 2015 in Review. https://www.eff.org/de/deeplinks/2015/12/lets-encrypt-project-comes-fruition-2015-review, 2015. Accessed: May 26, 2023.

[45] Richard Shay, Saranga Komanduri, Adam L. Durity, Phillip S. Huh, Michelle L. Mazurek, Sean M. Segreti, Blase Ur, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. Designing password policies for strength and usability. *ACM Transactions on Information and System Security (TISSEC)*, 18(4):13, May 2016.

[46] Richard Shay, Saranga Komanduri, Patrick G. Kelley, Pedro G. Leon, Michelle L. Mazurek, Lujo Bauer, Nicolas Christin, and Lorrie F. Cranor. Encountering Stronger Password Requirements: User Attitudes and Behaviors. In *Proceedings of the Sixth Symposium on Usable Privacy and Security (SOUPS '10)*, pages 1–20, New York, NY, USA, 2010. Association for Computing Machinery.

[47] Alexander Sotirov, Marc Stevens, Jacob Appelbaum, Arjen K Lenstra, David Molnar, Dag Arne Osvik, and Benne de Weger. MD5 considered harmful today Creating a rogue CA certificate. In *25th Annual Chaos Communication Congress*, number CONF, 2008.

[48] Nick Summers. Do you really need to change your password every 90 days? https://blog.1password.com/should-you-change-passwords-every-90-days/, 2022. Accessed: May 26, 2023.

[49] Joshua Tan, Lujo Bauer, Nicolas Christin, and Lorrie F. Cranor. Practical recommendations for stronger, more usable passwords combining minimum-strength, minimum-length, and blocklist requirements. In *Proceedings of the 2020 ACM*

SIGSAC Conference on Computer and Communications Security, CCS '20, page 1407–1426, New York, NY, USA, 2020. Association for Computing Machinery.

[50] Christian Tiefenau, Maximilian Häring, Katharina Krombholz, and Emanuel von Zezschwitz. Security, availability, and multiple information sources: Exploring update behavior of system administrators. In *Sixteenth Symposium on Usable Privacy and Security (SOUPS 2020)*, pages 239–258. USENIX Association, August 2020.

[51] Christian Tiefenau, Emanuel von Zezschwitz, Maximilian Häring, Katharina Krombholz, and Matthew Smith. A Usability Evaluation of Let's Encrypt and Certbot: Usable Security Done Right. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, CCS '19, page 1971–1988, New York, NY, USA, 2019. Association for Computing Machinery.

[52] Blase Ur, Fumiko Noma, Jonathan Bees, Sean M. Segreti, Richard Shay, Lujo Bauer, Nicolas Christin, and Lorrie F. Cranor. "I Added'!'at the End to Make It Secure": Observing Password Creation in the Lab. In *Eleventh Symposium On Usable Privacy and Security (SOUPS) 2015*, pages 123–140, Ottawa, July 2015. USENIX Association.

[53] W3C. W3C TAG: Securing the Web. https://www.w3.org/2001/tag/doc/web-https, 2015. Accessed: May 26, 2023.

[54] W3Techs. W3techs: Historical yearly trends in the usage statistics of site elements for websites. https://w3techs.com/technologies/history_overview/site_element/all/y, 2022. Accessed: May 26, 2023.

[55] Yinqian Zhang, Fabian Monrose, and Michael K. Reiter. The security of modern password expiration: An algorithmic framework and empirical analysis. In *Proceedings of the 17th ACM conference on Computer and communications security*, CCS '10, pages 176–186, New York, NY, USA, 2010. Association for Computing Machinery.

A Survey

The used survey was adapted to new situations over the years. To indicate that a question was included in a year, we will use the following taxonomy:

- †, if a question was part of the original questionnaire by Gerlitz et al. [22] in **2019**.
- ⋆, if the questions was included in **PCP20**.
- ◇ for questions included in **PCP22**.
- ▽ for questions included in **PCP23**.

If no year is specified, the question was asked in all versions of the survey.

Accounts

- Q1 Is there a company-wide account per user, that is managed centrally? (e.g., for logging into the workstation, communication platform, email or the like.)
  Yes / No
- *Q2*$^{\triangledown}$ Which user management do you use? (e.g. Microsoft Active Directory)
  [Free Text]

If Q1 equals yes:

- Q3a What can this account be used for? (Multiple answers possible.)
  Email / Workstations / Communication platform (SharePoint, Slack, etc.) / VPN into corporate network / Access to shared corporate data (e.g., Active Directory) / Other: [Free text]

< Page break >

- Q3b Which methods can be used to log in? Please check the applicable.
  Password or PIN / Biometrics (e.g., Fingerprint, Face recognition) / Hardware Token (e.g. Smartcard, Token, Smartphone) / Device Certificates$^{\diamond}$
- Q3c Is there any other method in use that is not listed?
  Yes, the following: [Free text] / No

If more than one method is listed in Q3b:

- Q4a You stated, that there are several methods in use that enable your employees to log in. Are the methods used in combination (e.g., 2FA)?
  Yes, the methods are used in combination (2FA) / No, the employees can choose one of the methods / Other: [Free text]/ I do not know / I do not wish to make a statement

If only one method is listed in Q3b:

- *Q4b*$^{\triangledown}$ Do you make use of any kind of two-factor authentication?
  Yes, using the following techniques: [Free Text] / No / I do not know / I do not wish to make a statement

< Page break >

Passwords

You stated that there is no company-wide account with which the employees can log into several services.
The following questions regard the email accounts of your employees and their passwords (Webmail, IMAP, POP3, etc.).
Or
You stated, that your employees use passwords/PINs to log in. The following questions regard these passwords/PINs.

- *Q5*$^{\dagger\star\diamond}$ How are passwords handled?
  *Users can choose them themselves / Passwords are created by a system, and users **cannot change** them / Passwords are created by a system and **need to be changed** by the user*$^{\star\diamond}$*/ I don't want to make a statement / Other: [Free text]*
- Q6 What specification (also called password policy) do passwords need to fulfill (e.g., at least x characters, new password needs to be selected after x days, etc.)

  This question is the main focus of our research. Please be as detailed as possible. If possible and allowed, please copy your specification into the following text box. At this point, we want to remind you, that the data is managed anonymously. It will not be possible to identify your company.
  [Free text]

- Q7 Are these specifications enforced by the system?
  Yes / No / There are no specifications / I do not know / I do not wish to make a statement / Partially: [Free text]
- $Q8^{\diamond}$ In case there is a password expiry in use: Why?
  [Free text]
- Q9 Optional: What reasons spoke against the introduction of a password policy?
  [Free text]

< Page break >

- $Q10^{\triangledown}$ Do you feel it is best practice to require employees to change their passwords on a regular basis (e.g., once a year)?
  Yes / No / Other: [Free Text]

< Page break >

In previous studies, we have seen that employees in many companies in Germany have to change their passwords on a regular basis. In the following, we would like to learn more about the possible reasons that speak in your eyes for or against using such a time-controlled change of passwords.

- $Q11^{\triangledown}$ Are employees technically forced to change their password?
  Yes, after a fixed time interval (days/months/years): [Free Text] / Yes, if there is a suspicion that the password has become known / No, never / Other: [Free Text]
- $Q12^{\triangledown}$ If employees need to change their password regularly (after a fixed time interval): Why?
  [Free Text]
- $Q13^{\triangledown}$ In case employees had to change their password regularly (after a fixed time interval) in the past, but this rule has now been abolished: Why was this changed?
  [Free Text]

< Page break >

- $Q14^{\triangledown}$ Since 2020, the BSI has recommended relying on password compromise detection instead of recurring password change prompts (after a fixed time interval). If you do this: How do you check if passwords are compromised?
  If not: are there reasons that prevent you from doing so? (e.g., technical, structural, or timing reasons).
  [Free Text]

< Page break >

- Q15 Are users prevented from picking passwords that belong to the most common passwords?
  No / Other: [Free text] / I do not know / I do not wish to make a statement / Yes[†] / Yes, examination through$^{\star\diamond}$: [Free text])
- $Q16^{\star\diamond\triangledown}$ Do you make use of a Blocklist/Denylist of elements that are not allowed to be used in a password? (e.g. words from the dictionary or sequences of numbers) *Yes, the following elements cannot be used in passwords: [Free text] / No / There are no specifications / I do not know / I do not wish to make a statement*
- $Q17^{\star\diamond}$ Which Unicode characters can be used in passwords?
  All / a-z / A-Z / 0-9 / All special characters / Special characters, except: [Free text] / Chinese characters / Arabic characters / Emojis / Other: [Free text] / I don't know / I do not want to make a statement
- $Q18^{\star\diamond\triangledown}$ Are there additional password policies for different users? (e.g. System administrators) *Yes / No / I do not know / I do not wish to make a statement*
- $Q19^{\star\diamond\triangledown}$ Optional: How do the different password policies differ? *[Free text]*

< Page break >

The following questions still regard the passwords which are used in your company.

- Q20 Who created the specifications (password policies) for the passwords?
  Myself / My predecessor / Somebody else: [Free text]/ I do not know / I do not wish to make a statement / There are no specifications
- Q21 What are the specifications based on? (Multiple answers possible.)
  Targeted Training$^{\star\diamond\triangledown}$ / Own Know – how$^{\star\diamond\triangledown}$ / Standards defined by own company$^{\star\diamond\triangledown}$ / Industrial Standards$^{\star\diamond\triangledown}$ / Expert panels / Exchange with other companies / NIST (National Institute of Standards and Technology) / BSI (Bundesamt für Sicherheit in der Informationstechnik) / OWASP (Open Web Application Security Project) / Other: [Free text] / I do not know / I do not wish to make a statement
- $Q22^{\star\diamond\triangledown}$ When were the password policies changed last? *[Free text]*
- $Q23^{\star\diamond}$ What changes were made during the last modification?
  [Free text]
- $Q24^{\triangledown}$ Which changes were made?
  [Free Text]
- $Q25^{\star\diamond\triangledown}$ What caused the change? *[Free text]*
- Q26 How do the password policies impact the user-friendliness of the authentication system?
  1: Very negative – 5: Very positive
- Q27 How do the password policies impact the security of the authentication system?
  1: Very negative – 5: Very positive
- Q28 How often do passwords cause problems in your company (e.g., forgotten passwords, etc.)?
  1: Very rarely – 5: Very often

< Page break >

- $Q29^{\star\diamond}$ This year, the BSI published new recommendations for password policies. Have you already dealt with them? *Yes / No / I do not know / I do not wish to make a statement*
- $Q30^{\star\diamond}$ Was your password policy adapted due to the new recommendations or are you planning a change? *Yes / No / I do not know / I do not wish to make a statement*

< Page break >

- Q31 Is there a policy that specifies how the passwords are stored in the system (hash function, length of the salt, etc)?
  Yes / No / I do not know / I do not wish to make a statement
- Q32 Is there a process that initiates an update of the policy on how to store passwords?
  Yes / No / I do not know / I do not wish to make a statement

- Q33 Optional: How are stored passwords protected? We are particularly interested in the hash and salt functions that are used.

  We want to remind you that the data is gathered anonymously and we are not able to link it to your company.
  [Free text]

< Page break >

Biometric Authentication

You stated, that your employees use biometrics to log in. The following questions regard this method.

- Q34 What kind of biometrics are in use?
  Fingerprint / Iris / Face recognition / Other: [Free text]/ I do not wish to make a statement
- Q35 How does the biometric authentication impact the user-friendliness of the authentication system?
  1: Very negative – 5: Very positive
- Q36 How does the biometric authentication impact the security of the authentication system?
  1: Very negative – 5: Very positive
- Q37 How often does the use of biometric authentication cause problems?
  1: Very rarely – 5: Very often
- Q38 Optional: Do you wish to provide us with additional information about this topic?
  [Free text]

< Page break >

Hardware Token

You stated, that your employees use a hardware token to authenticate. The following questions regard this token.

- Q39 Does the token support FIDO2?
  Yes / No / I am not sure / I do not wish to make a statement
- Q40 How does the token impact the user-friendliness of the authentication system?
  1: Very negative – 5: Very positive
- Q41 How does the token impact the security of the authentication system?
  1: Very negative – 5: Very positive
- Q42 How often does the usage of the token cause problems?
  1: Very rarely – 5: Very often
- Q43 Optional: Do you wish to provide us with additional information about this topic?
  [Free text]

< Page break >

Passwordless

- $Q44^{\triangledown}$ There are efforts (e.g., by the Fido Alliance) to completely abolish passwords. What is your opinion of these efforts (also with regard to their feasibility in the company where you work)?
  [Free Text]

< Page break >

Demographics

- $Q45^{\dagger\star\diamond}$ Please check the conditions which apply to your company. (Multiple answers possible.)
  There are employees who can access their emails outside the company network /
  There are employees who can access their emails using a web login /
  There are employees who do not need to know the password for accessing their emails, e.g., as the email-client is pre-configured
- Q46 Is there any additional security for emails? (e.g., encryption in combination with a smart card)
  Yes, obligatory / Yes, voluntary / No / I do not wish to make a statement
- $Q47^{\star\diamond\triangledown}$ What is your companies' field of work? *Automobile Industry / Banks and financial services / Education and research / Services / Retail / Energy industry / Logistics / Telecommunication / Pharmaceutical industry / Tourism / Insurance / Healthcare Marketplace / Other: [Free text] / I do not wish to make a statement*
- $Q48^{\star\diamond\triangledown}$ Is your company an operator of critical infrastructure or is it affected by regulations for operators of critical infrastructure? *Yes / No / I do not know / I do not wish to make a statement*
- $Q49^{\triangledown}$ Is your company certified with a certification relevant to IT security (e.g., PCI-DSS, ISO 27001,..)?
  Yes, the following: [Free Text] / No, but we intend to / No / Other: [Free Text] / I don't know / I do not wish to make a statement
- $Q50^{\triangledown}$ Are there any legal regulations that require you to have any of the previous certifications?
  Yes the following: [Free Text] / No / Other: [Free Text] / I do not wish to make a statement
- $Q51^{\star\diamond\triangledown}$ Where is your companies' headquarter? (Country)
  [Free text]
- Q52 How many employees work in your company?
  1-9 / 10-49 / 50-249 / 250-499 / 500-999 / 1000 or more / Not sure / I do not wish to make a statement
- Q53 How many desktop clients do you manage?
  1-9 / 10-49 / 50-249 / 250-499 / 500-999 / 1000 or more / Not sure / I do not wish to make a statement
- Q54 How many employees in your company work full-time on IT security topics?
  0 / 1 / 2-5 / 6-10 / 11-20 / 21 or more / Not sure / I do not wish to make a statement

- $Q55^{\star\diamond\triangledown}$ What is your position? *Administrator / ISO / CISO / CTO / CSO / Support / Other: [Free text] / I do not wish to make a statement*
- $Q56^{\star\diamond\triangledown}$ How many years of experience do you have in this or related positions? *Under 1 year / 1-3 Years / 4-9 Years / 10 or more years / I do not wish to make a statement*
- $Q57^{\star\diamond\triangledown}$ Is one (or more) of the following situations true for your company when looking at the last 5 years? (MC) *A password of an employee was guessed and used to attack the company (Ransomware, theft,..) / Several passwords have been stolen from the database / None of the above / I do not know / I do not wish to make a statement / Further / Other: [Free text]*

`< Page break >`

- $Q58^{\star\diamond\triangledown}$ Have you already participated in this survey last year? *Yes / No / I do not know /I do not wish to make a statement*
- Q59 How satisfied are you with your authentication system? *1: ☹ – 5: ☺*
- Q60 Has this questionnaire motivated you to update parts of your authentication system in the near future? If yes, which parts?
  Password Policies / Security measures for stored passwords / Adding biometrics / Adding hardware token / No / Other: [Free text]

B Additional Tables

	PCP19	**PCP20**	**PCP22**	**PCP23**
Pw	100.0	100.0	100.0	97.5
Pw + Bio	14.8	17.3	25.4	25.0
Pw + Token	38.9	50.0	57.1	53.8
Pw + Bio + Token	11.1	13.5	19.0	17.5
Token (no Pw + Bio)	0	0	0	1.2

Table 2: Percentage of participants who mentioned that their companies use the different possibilities to login. Pw = Passwords are in use; bio = Biometrics are in use; Token = Hardware token are in use

	PCP19	**PCP20**	**PCP22**	**PCP23**
All Data	172	72	96	122
Only BSI	91	72	96	122
Only complete	71	57	66	83
Individual filter	69	56	66	83
Only accounts	**54**	**52**	**63**	**80**

Table 3: Elimination process of data sets. **Only BSI** = Only answers that were gathered over the BSI newsletter. Only in PCP19 were participants recruited over other channels as well. **Complete** = The participant filled out the whole survey. **Individual filter** = Participants were filtered manually if the answers indicated they did not understand the questions and if the role did not include being able to work on the company's authentication protocols. **Only accounts** = Participants indicated whether the company makes use of a centrally managed account. We only kept those who did.

		PCP19	PCP20	PCP22	PCP23
		$n = 54$	$n = 52$	$n = 63$	$n = 80$
Size of Company (Q52)	1-9	7.4	7.7	7.9	5.0
	10-49	13.0	3.8	11.1	13.8
	50-249	16.7	23.1	15.9	20.0
	250-499	7.4	7.7	9.5	12.5
	500-999	9.3	9.6	11.1	16.2
	≥ 1000	46.3	48.1	44.4	28.7
	Unclear	0.0	0.0	0.0	3.8
Employees working full-time on IT security topics (Q53)	0	14.8	5.8	17.5	17.5
	1	22.2	36.5	27.0	23.8
	2-5	25.9	34.6	31.7	31.2
	6-10	5.6	11.5	7.9	8.8
	11-20	11.1	5.8	0	7.5
	≥ 21	13.0	5.8	11.1	7.5
	Unclear	7.4	0.0	4.8	3.8

Table 4: Demographics of companies that were asked in all three years. All numbers are percentages of that year. "Unclear": Participants did not disclose the information, or we could not infer it from their answers.

		PCP20 $n=52$	PCP22 $n=63$	PCP23 $n=80$
Sector (Q47)	Services	40.4	39.7	41.2
	Industry	17.3	15.9	15.0
	Medical	7.7	7.9	5.0
	Infrastructure	9.6	4.8	7.5
	Public service	9.6	9.5	6.2
	Education and research	5.8	9.5	6.2
	Sales	3.8	1.6	3.8
	Other	0.0	7.9	5.1
	n.d.	5.8	3.2	10.0
Critical Infrastructure (Q48)	Yes	19.2	15.9	13.8
	No	80.8	74.6	81.2
	n.d.	0	9.5	5.0
Incidents at company within last 5 years (Q57, MC)	Easy PW used for attack	11.5	9.5	11.2
	Several PWs were stolen from database	1.9	1.6	0.0
	None of the above	63.5	84.1	65.0
	Other	11.5	0.0	8.8
	n.d./Don't know	15.4	4.8	15.0
Own role (Q55)	IT			
	C-level & management	50.0	39.7	50.0
	ISO	17.3	28.6	20.0
	Admin/DevOps	13.5	12.7	11.2
	Support	0	1.6	0
	Management	1.9	9.5	7.5
	DPO	1.9	0.0	2.5
	Consultant	0.0	0.0	1.2
	n.d.	17.3	7.9	7.5
Own experience (Q56)	<1 year	1.9	1.6	2.5
	1-3 years	17.3	17.5	12.5
	4-9 years	30.8	30.2	28.7
	$\geq$10 years	44.2	50.8	52.5
	n.d.	5.8	0	3.8

Table 5: Demographics of companies and participants from PCP20, PCP22, and PCP23. All numbers are percentages of that year. "n.d": Participants did not disclose their answers. The participants indicated their current job position. Since some of them indicated holding different roles in the company (e.g., CEO and admin), the numbers exceed 100%.

	Code	PCP20 $n=52$	PCP22 $n=63$	PCP23 $n=80$	ICR
Character classes	1	2	1	1	-
	2	2	3	1	0.79
	3	23	25	33	1
	4	16	19	15	1
	Imprecise	3	5	9	0.47
	Special	0	1	2	-
	Total	46 (88%)	54 (86%)	61 (76%)	
	Not mentioned	4	6	10	0.79
	Explicitly not	0	0	1	-
Minimum length	5	1	0	0	-
	6	4	1	1	-
	7	0	0	1	-
	8	18	20	19	1
	9	1	1	2	1
	10	16	14	12	1
	11	0	2	1	1
	12	8	16	24	1
	14	0	2	5	1
	15	0	1	4	-
	16	1	0	2	-
	64	1	0	0	-
	Imprecise	0	2	1	-
	Total	50 (96%)	59 (94%)	72 (90%)	
	Not mentioned	0	1	0	-
Password expiry (days)	14	0	1	0	-
	30	1	0	2	-
	42	1	1	0	1
	60	0	0	1	-
	64	0	0	1	-
	90	12	8	5	1
	120	0	0	1	-
	168	0	1	0	-
	180	6	2	9	1
	230	0	1	0	-
	360	0	0	1	-
	365	5	6	6	1
	540	1	0	0	-
	720	0	0	1	-
	Imprecise	1	2	1	-
	Total	27 (52%)	22 (35%)	28 (35%)	
	Not mentioned	18	25	26	1
	Explicitly not	5	13	18	1
	No policy given	2	3	8	-

Table 6: Codebook of participants' elements of password composition policies (Q6) and how often they occurred. For calculating the ICR, answers from both years were merged. Some codes were not covered in the documents that were used to calculate the inter-coder reliability and indicated as "-" in the table.

Code	Example	Occurrence in PCP23
Demanded by ...		
... Institutions	"PCI DSS (Credit card security) requirement"	5
... Customer (contracts)	"was embedded in customer contracts in the past"	3
... Own company	"In-house or internal definition"	4
Is best practice	"Recommendation by BSI"	2
To increase security	"Improve password security", "Ensure that there are no passwords that are too old and no insecure hash algorithms are used."	17
Currently changing	"We still have the setting but are in discussion rather to increase the complexity, but not to force a change (except in case of loss). Still need to convince the auditor. "	1
Alternative not implemented	"Because there is currently no other technical solution."	4
Inertia	"Still exists from history"	5

Table 7: Codebook of the reason for using password expiry (Q12)

Code	Example	Occurrence in PCP23
Employee check	"Our detection of compromise has so far been the sole responsibility of the employee."	6
Check: Dark web monitoring	"Dark Web Monitoring"	4
Check: Compare to leak database	"On the relevant websites we check whether passwords have been compromised"	8
Check: Anomaly detection	"Detection of compromise by technical means (e.g., logon location)."	16
No check: Lack of..		
..technical measures	"Technically not yet possible, corresponding recognition systems are still missing"	9
...and unclear how	"We do not know any practicable method"	3
...organizational measures	"organizational feasibility"	5
...resources (time, money)	"There would also be a lack of time and staff resources to look into this"	7

Table 8: Codebook of the reason for using password expiry (Q14)

		Has PW Expiry	No PW Expiry	p	OR
H3: Critical Infrastructure	Yes No	5 28	3 14	1.0	0.83
H4: Company Size	<500 >=500	14 16	9 8	1.0	1.29
H5: Last policy change	Before BSI changes After BSI changes	9 19	1 15	0.3	0.14
H6: Technical compromise check	Yes No	13 20	12 5	0.4	0.27

Table 9: Contingency table for H3-6: *Company characteristics or the use of certain policy elements influence whether the companies use a password expiry in 2023.* The results are corrected using a Bonferroni-Holm correction.

Dissecting Nudges in Password Managers: Simple Defaults are Powerful

Samira Zibaei
samira.zibaei@ontariotechu.net
Ontario Tech University

Amirali Salehi-Abari
abari@ontariotechu.ca
Ontario Tech University

Julie Thorpe
julie.thorpe@ontariotechu.ca
Ontario Tech University

Abstract

Password managers offer a feature to randomly generate a new password for the user. Despite improving account security, randomly generated passwords (RGPs) are underutilized. Many password managers employ *nudges* to encourage users to select a randomly generated password, but the most effective nudge design is unclear. Recent work has suggested that Safari's built-in password manager nudge might be more effective in encouraging RGP adoption than that of other browsers. However, it remains unclear what makes it more effective, and even whether this result can be attributed to Safari's nudge design or simply Safari users. We report on a detailed large-scale study of Chrome users (n=853) aimed at clarifying these issues. Our results support that Safari's nudge design remains more effective than Chrome's among Chrome users. By dissecting the elements of Safari's nudge, we find that its most important element is its *default* nudge. We additionally examine whether a social influence nudge can further enhance the Safari nudge's RGP adoption rate. Finally, we analyze and discuss the importance of a nudge being noticed by users, and its ethical considerations. Our results inform RGP nudge designs in password managers and should also be of interest to practitioners and researchers working on other types of security nudges.

1 Introduction

Passwords remain the most widely-deployed form of authentication over the Web. Users are expected to manage and remember many passwords for many web services and accounts. To cope with remembering numerous passwords, users resort to reusing passwords across different accounts, leaving them vulnerable to credential stuffing attacks [20, 43]. Password managers are one solution to this problem, as long as users make use of their feature to generate and save a random and unique password for each account [30]. Unfortunately, these *randomly generated passwords (RGPs)* are infrequently used (e.g., only 35% of Chrome users [51]). As such, encouraging the use of RGPs in password managers is an important method to improve users' online security [27, 32]. Popular web browsers (e.g., Chrome, Firefox, Safari, etc.) have encouraged the use of RGPs at the time of password creation through *nudging*, without limiting user choices such as typing their own passwords [1–3].[1] A recent study found that Safari users are more likely to adopt an RGP than Chrome or Firefox users [51], suggesting that the underlying cause might be Safari's nudge design. Despite being interesting, this finding has raised many unanswered questions when it comes to the adoption of RGPs: **(Q1)** Does Safari's nudge design remain more effective among users of another browser (e.g., Chrome) or was it just that Safari users were more inclined to adopt RGPs? **(Q2)** Which design components of Safari's nudge (e.g., nudge types it employs) contribute to its high RGP adoption rate? **(Q3)** Can we further extend Safari's RGP adoption rate by incorporating other promising nudging techniques (e.g., social influence[2])?

We evaluate these questions within a population of Chrome users through a large-scale online study (n=853). Participants were asked to register for an account to test a new e-commerce website, during which we observed their interaction with a browser-based password manager. We focus on browser-based password managers to avoid overhead and issues with installing standalone password managers. Each participant

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.
USENIX Symposium on Usable Privacy and Security (SOUPS) 2023.
August 6–8, 2023, Anaheim, CA, United States.

[1] Nudging can broadly be defined as shaping the choice environment to encourage the adoption of a specific choice over others, while not limiting the possible choices [26].

[2] *Social influence nudges*, by providing descriptive information on other people's actions, give the impression that the desired action is approved and accepted by other people [10].

was assigned to one of six nudge design conditions described in Section 3. We perform quantitative analyses on our server logs to find which nudge design works best in terms of RGP adoption rates. To understand users' reasons for adoption (or rejection) of RGPs, we perform a qualitative analysis of users' provided reasons for RGP adoption or rejection. Our key contributions are:

- We provide evidence that Safari's nudge design is indeed more effective than Chrome's. Zibaei et al. [51] only tested Safari's nudge on Safari users; we show the result holds for Chrome users as well.
- We provide the first evaluation of the efficacy of Safari's nudge design elements. Our findings reveal that the default nudge, which automatically populates the password field with a RGP, is the most important element.
- We find that our specific social influence nudge design did not significantly improve the efficacy of Safari's nudge.
- We explore factors that might explain or contribute to the efficacy of nudges in this study. Our findings confirm the results of Zibaei et al. [51], in that they suggest that prior experience using RGPs can have a significant impact on users adopting RGPs.
- We perform a qualitative analysis of users' reasons and barriers to RGP adoption. Surprisingly, security concerns are reasons for and against RGP adoption.

In addition to these contributions, our work sheds light on the importance of whether a nudge is noticed by users. In particular, our results suggest that noticeability is only important for the efficacy of some nudge designs, and that many designs do not take advantage of the attention they draw. We discuss users' reasons for still not adopting the RGP even when they noticed the nudge. Our work provides strong evidence for the effectiveness of default nudges to encourage more secure user behaviors. We further discuss the ethics of default nudges and argue that noticeability can be an important design goal from an ethical perspective.

The remainder of this paper is organized as follows. Section 2 describes related work. The methodology including implementation details and nudge designs is described in Section 3. We report and analyze the results in Section 4. We discuss our findings in Section 5. Concluding remarks and future work are discussed in Section 6.

2 Related Work

Password managers are critical solutions for storing and suggesting secure passwords to users over the Web. However, they are not widely-adopted or at least not used to their full potential [32, 33], partially due to some security and usability concerns. We review the research findings on why password managers are (not) adopted, relevant improvements proposed for password managers, and relevant research on nudging.

2.1 Why Adopt Password Managers?

A growing body of research has focused on understanding which characteristics of password managers and their users contribute to their adoption. Ease of use has been reported as the primary reason for password manager adoption [29]; this can relate to the save password feature [7], auto-fill, user interface design, and ease of installation process [40]. Other reported important features include reliable encryption methods and secure cloud backups of the passwords [28]. While there may be different reasons to adopt password managers for different user groups (e.g., based on gender [49]), in general, it appears that cybersecurity knowledge plays an important role in the adoption of password managers [24, 25].

2.2 Barriers and Improvements

Numerous studies have examined user's barriers to adopting password managers. A variety of obstacles have been identified, including lack of awareness [5, 32, 39], lack of knowledge [5], complex user interface and terminology [5, 40, 41], and lack of trust and transparency [5, 7, 15, 22, 33]. Security concerns such as the risks of a single point of failure [32] and unauthorized access to stored passwords [33] have also hindered adoption. Some users also believe they do not have many accounts to be worried about [32]. Additional reasons for rejecting password managers include the burden of installing standalone software and perceiving them as unnecessary security tools [7, 8, 11].

Efforts to address these adoption barriers have focused on minimizing users' required actions [42], improving password manager user interfaces [9,42], recommending password managers tailored to user requirements [6], introducing educational videos [38], and addressing password reuse issues [41].

2.3 Nudging

Nudging is a behavioral and decision-making technique that influences people's choices without mandating a specific outcome. Nudging theory has been employed in a wide range of human-computer interaction [10] and cybersecurity [52] topics. Some notable examples in cybersecurity and privacy applications involve joining a safe Wi-Fi network [50], making a social post [48], trusting received emails [13], password meters [14, 16], and the problem of password creation in alphanumerical passwords [36, 37, 45, 47] and graphical passwords [31, 44]. Default nudges leverage individuals' tendency to choose the default option instead of considering other alternatives; they have been shown to have a significant effect on changing user behavior across a wide variety of domains [17, 19]. Recently, it has been shown that some popular

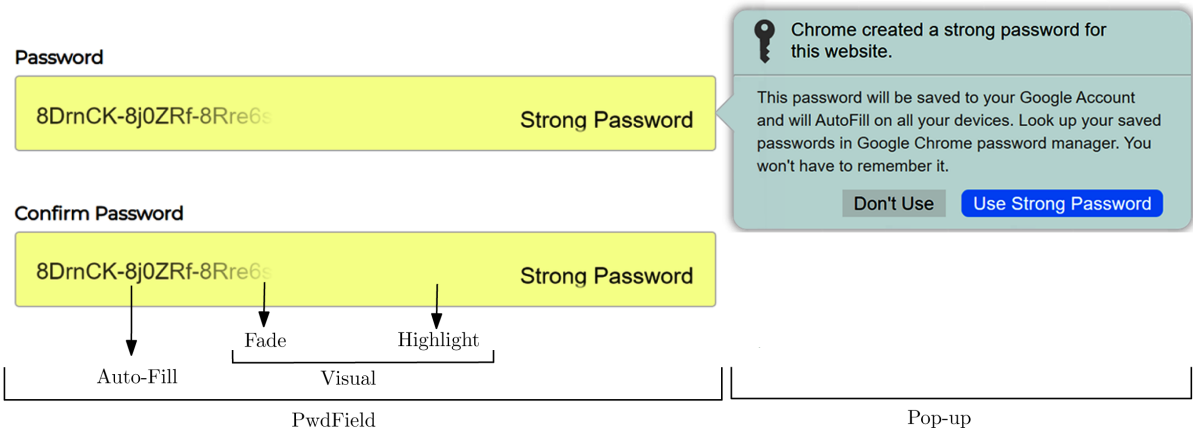


Figure 1: Dissection of Safari's nudge elements: (Left) PwdField is part of the nudge that modifies the password input field, consisting of both Auto-fill and Visual elements. The visual elements are fading the tail of the auto-filled password to make it appear longer and highlighting the password field in yellow. (Right) Pop-up provides a detailed message informing users about the randomly generated password, where it is stored, and reinforcing that they won't need to remember it.

web-browser-based password managers are more effective than others in motivating their users to adopt RGPs, with Safari being the most effective [51].

Our work not only replicates recent findings on password manager nudging [51], but also extends them by answering several critical unanswered questions. Specifically, we investigate whether the high RGP adoption rate of Safari is due to Safari's nudge design or its users. Through a detailed analysis of Safari's UI design elements, we identify and study the specific components that contribute to its high RGP adoption rate. Additionally, we propose and evaluate a social influence nudge enhancement for Safari, which has the potential to further increase RGP adoption rates.

Table 1: Nudge types employed in UI design elements.

UI elements	Default	Social Comparisons	Deceptive Visual	Suggest Alternatives	Just-in-time Prompt
Chrome's UI				✓	✓
Autofill	✓			✓	✓
Pop-up				✓	✓
Visual			✓		✓
Social Pop-up		✓		✓	✓

3 Methodology

Our main goal is to independently evaluate the efficacy of Safari's UI elements in nudging RGP adoption. Our secondary goal is to evaluate the effectiveness of social nudges as an extension of Safari's nudge. To also reproduce findings that suggest Safari is the most effective in nudging RGPs [51], we used a similar methodology and implementation as Zibaei et al. [51][3] but with a population of Chrome users only. We collect quantitative data by collecting user interactions with the password manager, and a post-study questionnaire where participants were asked to answer some questions regarding their actual behavior and intentions. Our study was reviewed and approved by our institution's Research Ethics Board.

[3] We have extended their implementation found at https://github.com/rinoa25/Secure-Password-Creation-Nudge-Prototypes.

3.1 Dissecting Safari's Nudge

We implemented a version of Safari's nudge design (see Figure 1) for Chrome, in order to compare its efficacy to Chrome's nudge between samples of Chrome users. Safari's nudge design can be broken down into various UI elements as labeled in Figure 1. At the highest level, it has two main UI elements: the password field (*PwdField*) and the message box (*Pop-up*). The password field can be further broken down into two UI elements: an auto-filled RGP (*Auto-Fill*) and visual effects (*Visual*). The visual effects *highlight* the password field to make it more visually striking and *fade* the last 6 characters of the suggested password to give the impression of a long password with many characters.

The message box in Safari normally has the heading of *"Safari created a strong password for this website"* and the message of *"This password will be saved to your iCloud Keychain and will AutoFill on all your devices. Look up your saved passwords in Safari Password preferences or by asking*

Table 2: UI design elements for non-Chrome conditions.

Conditions	Safari's UI			Social
	Autofill	Visual	Pop-up	Pop-up
Safari	✓	✓	✓	
PwdField	✓	✓		
PwdField-No-Visual	✓			
Pop-up			✓	
Safari-Social	✓	✓	✓	✓

Siri." This message emphasizes convenience by *"AutoFill"* and explains the storage place by *"iCloud Keychain"*, aiming to educate users on the password manager's functionality. The security is highlighted by a *"Use Strong Password"* button that users must select if they wish to accept the RGP. In our implementation, we have slightly reworded the pop-up message to be consistent with the fact it is running on Chrome (and as such would be saved to the Google Account and can be looked up on the Google Chrome password manager).

The UI elements of Safari as discussed here, as well as the UI of Chrome, employ various types of nudges as described in Table 1 and Section 3.3.

3.2 Prototypes and Conditions

We use a between-groups study design where each group was in a separate condition that used one of the following prototypes. Our prototypes were implemented on the Chrome browser as it is the most popular web browser [4]. Many of the prototypes implement a subset of the UI elements identified in our dissection of Safari's nudge (recall Section 3.1 and Figure 1). The full list of conditions (or prototypes) and their UI elements are summarized in Table 2 and described below:

- The *Safari* prototype simulates Safari's nudge design and interface on Chrome with a minor difference in the wording of the pop-up text to customize it to Chrome: *"Chrome created a strong password for this website. This password will be saved to your Google Account and will AutoFill on all your devices. Look up your saved passwords in Google Chrome password manager. You won't have to remember it"* (see Figure 1). This prototype includes the Autofill, Pop-up, and Visual UI elements.
- The *PwdField* prototype simulates only the PwdField part of Safari's nudge interface. It retains Chrome's informative messaging "Chrome will save this strong password in your Google Account. You won't have to remember it." (see Figure 2). This prototype contains the Autofill and Visual UI elements.
- The *PwdField-No-Visual* prototype simulates only the Autofill UI element of Safari's nudge interface without the visual effects (see Figure 3).

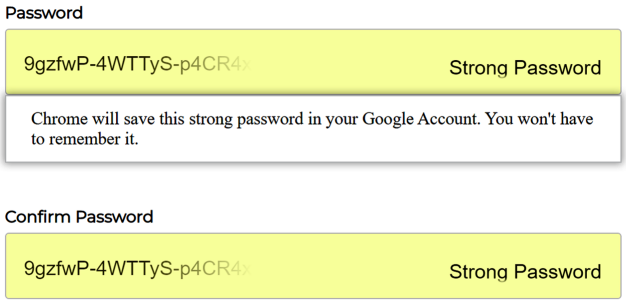


Figure 2: PwdField prototype simulates only the PwdField part of Safari's nudge interface.

- The *Pop-up* prototype simulates only the Pop-up UI element of Safari's nudge interface (see Figure 4).
- The *Safari-Social* prototype (see Figure 5) enhances the Safari prototype with a social influence nudge. This prototype is motivated by the success of social influence nudges in other contexts (e.g., online shopping) [17, 18, 46]. We chose to investigate a social nudge due to its promising results across a variety of fields [17, 18], and also its absence in Safari's nudge design. This prototype includes the Autofill, Pop-up, and Visual UI elements from Safari's nudge, and modifies the Pop-up to include a social nudge in its text and buttons. The Social Pop-up is a pop-up containing the message: "Chrome created a strong password that will be saved and remembered for you", followed by "*Join other users and be part of the secure movement by using this strong password.* [emphasis added] You can look up the saved password in your Google Chrome password manager. You won't have to remember it." The italicized text intends to influence the user by giving the impression that accepting a strong random password is an approved and acceptable action by many other users. We modify the button labels within the pop-up message to reinforce that adopting the RGP is a desirable behavior.
- The *Chrome* prototype simulates Chrome's nudge and interface (see Figure 6). Key phrases of the Chrome popup also exist in our Pop-up prototype: "Use strong password", "You won't have to remember this password", and "it will be saved". Both contain key icons. The main difference is that Chrome's popup contains the RGP within it. Using Caraban's nudging classification [10], both the Pop-up and Chrome prototypes employ facilitate (suggesting alternatives) and reinforce (just-in-time prompt) nudges.

In our study, each prototype discussed above corresponds to a condition. We most often are interested in measuring

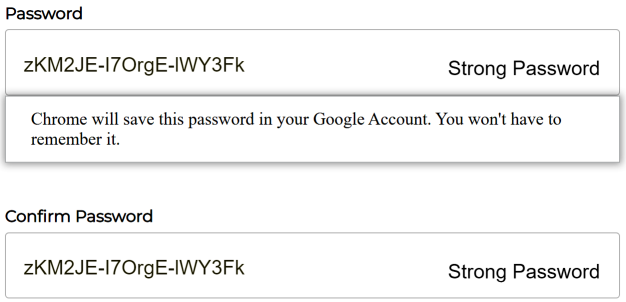


Figure 3: PwdField-No-Visual prototype simulates the PwdField part of Safari's nudge interface, without the Visual effects.

how the RPG adoption rate changes for various conditions with different nudging. We describe how each UI element implements different nudge types next in Section 3.3.

3.3 Nudge Types

We map each UI design element reported in Table 2 to a set of nudge types; see Table 1 for our mapping. Here we describe each of the nudge types and how each design element of Safari employs them:

- *Default* nudges leverage individuals' tendency to choose the default pre-selected option among many other alternatives. By pre-selecting a default option, decision-makers can influence the choices of individuals without restricting their freedom of choice [10]. The Autofill design element employs a default nudge by automatically filling an RGP in the password field.
- *Enabling social comparisons* refers to an individual's tendency to emulate other's behavior. This tendency compels individuals to take heed of the behavior of others and seek social validation when they experience uncertainty in their decision-making [12]. Our Social Pop-up design element implements this through its message "Join other users and be part of the secure movement by using this strong password", intended to evoke a sense that other people are accepting the RGP.
- *Deceptive Visualization* refers to making information relating to desired behaviors more prominent through visual illusions. The goal is to make individuals to focus on a visually-striking option, even if it is not necessarily the best choice [21]. Safari's Visual design element incorporates a fading effect on the characters of the autofilled RGP, creating the illusion that the password is longer than it actually is. Additionally, the Visual design element draws attention to this effect by highlighting the password field in yellow.
- *Suggesting alternatives* bring individuals' attention to the presence of options that may have been overlooked [10]. Except the Visual design element, all other elements in Table 1 employ this nudge type by suggesting the option of selecting an RGP. Chrome provides a small box below the password field that suggests and displays a RGP. Similarly, Autofill automatically fills the password field with a RGP. Pop-up and Social Pop-up suggest using an RGP in a pop-up message. Importantly, none of these design elements force users to choose these alternatives; they are merely presented as options.
- *Just-in-time prompts* seek to grab individuals' attention at the proper time [10]. In the case of encouraging RGP use, the proper time is the time of password creation. Each design element in Table 1 employs a just-in-time nudge by presenting the RGP option immediately after the user clicks on a password field, which is the moment when they are about to create a password.

3.4 Study Structure and Tasks

Participants were randomly assigned to one of the conditions (corresponding to the prototypes discussed in Section 3.2) and were only permitted to complete the study once. Our study, as with Zibaei et al. [51], is structured around six tasks in the following order:

1. *Initial deceptive consent*: To avoid unrealistic, biased user behavior that may draw unrealistic attention to the RGP nudge, we employ a deceptive consent form by deceptively declaring that our study aims at usability testing of an e-commerce website's registration. The users were asked to read and agree to this consent if they wish to continue.
2. *Account registration*: Participants were asked to register on our website, using their email address and password. Participants had the freedom to either create their own password or use a randomly generated password. This is where the nudge design is encountered.
3. *Demographic questionnaire*: Participants were required to answer five demographic questions, where they have the option of "prefer not to answer" for all questions.
4. *Login*: Participants were asked to log in to their accounts using their email address and password.
5. *Post-study questionnaire*: Participants were asked some questions regarding their behavior toward nudges, with an option of "prefer not to answer" for all questions.
6. *Debriefing*: Participants were debriefed about the actual purpose of the study (i.e., nudging) before ending their session, where they were asked to read carefully and agree if they wish to submit their data.

We used the same questionnaires as Zibaei et al. [51], including a post-study attention check question as a means of

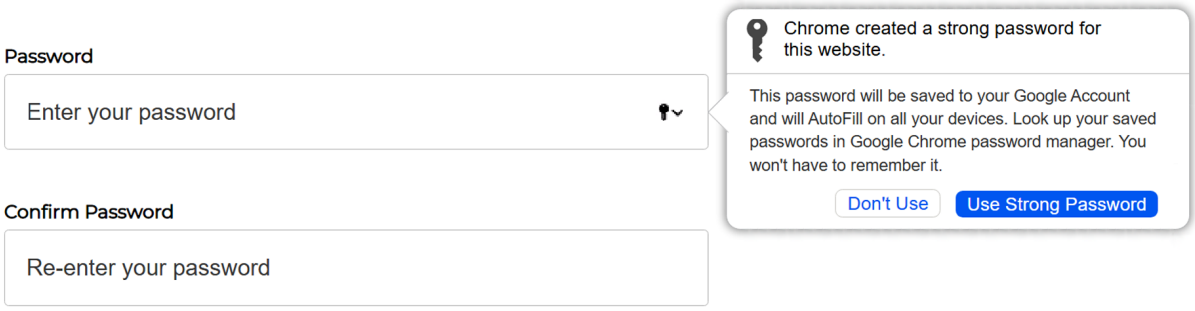


Figure 4: The Pop-up prototype simulates only the Pop-up UI element in Safari's nudge interface.

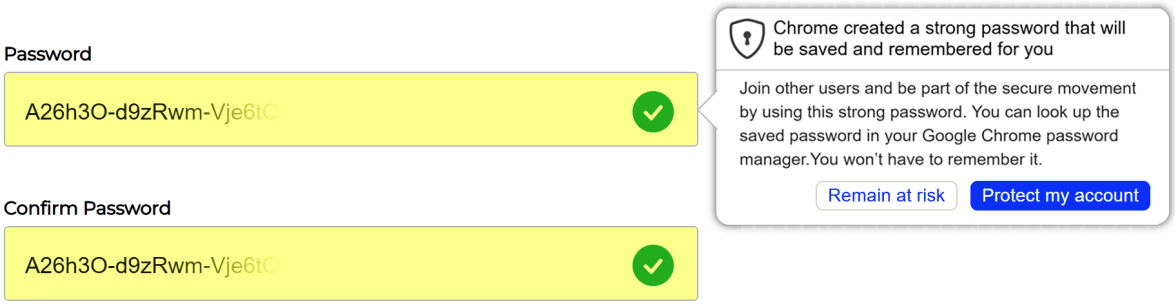


Figure 5: The Safari-Social prototype employs a combination of all nudge types in Table 1. It automatically fills a randomly generated password when the user clicks on the password field, grabs the user's attention by creating the illusion of suggesting a longer password, and suggests a randomly generated password as an alternative with additional social information in a pop-up message.

detecting poor quality data from inattentive participants. In order to minimize potential biases resulting from the questionnaires, we employed recommended guidelines [34].

3.5 Implementation Details

The UIs for each condition are simulated to remotely capture user interactions, while appearing to the user as the browser's PM. In order to achieve this, our website implements the UIs, and we don't define the password field as a proper password input to prevent the browser's PM from being invoked/interfering. The users were not aware of this simulation until the end of the study when we revealed it in a debriefing task. We record user interactions with the simulated password manager while creating an account and logging into the website. To ensure the confidentiality and quality of our collected data, we take a few key measures: (a) we ensure the actual built-in password manager of Chrome is not invoked for the account creation and login process; (b) we only collect passwords with anonymous identifiers, and we do not collect email addresses; (c) we only collect data once users have submitted the final debriefing form; and (d) users were allowed to participate only once in our study. As the participants of our study were expected to use Chrome as a web browser, we employed the user-agent header to confirm the correct browser was in use.

3.6 Recruitment

For our study, we recruited 862 participants via Amazon's Mechanical Turk platform, restricted to individuals from the United States. The duration to complete all study tasks was estimated to be less than 5 minutes. In accordance with the minimum wage in the United States ($7.25 USD per hour), participants were compensated $0.60 USD for their participation.

3.7 Analysis Method

We conduct both quantitative and qualitative analyses of our collected data. Our analyses aim to determine the effectiveness of Safari's nudge design and identify the key nudging elements that contribute to its effectiveness in encouraging RPG adoption. Specifically, we seek to determine whether there are statistically significant differences in RPG adoption

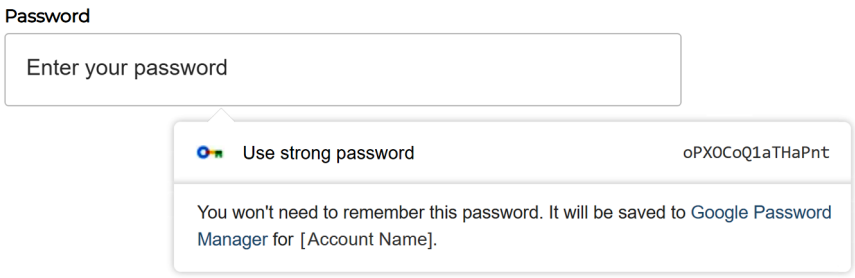


Figure 6: Chrome prototype simulates Chrome's nudge and interface (since Chrome 105, released Aug. 2022).

rates across the various condition groups, where all users in each group are exposed to a specific prototype introduced in Section 3.2. We use the χ^2 test, a statistical method for comparing proportions, to determine the statistical significance of adoption rates. For two condition groups of A and B (e.g., Safari vs. Chrome), our null and alternative hypotheses are:

H_0 The randomly generated password adoption rates are similar for two condition groups A and B.

H_a The randomly generated password adoption rates differ between two condition groups A and B.

We also use Bonferroni multi-test correction, and we report a result as statistically significant when it is significant after correction.

For the qualitative analysis, we assessed participants' comments in a post-study questionnaire, including an open-ended question that asked the participants to explain their reasons for (not) using an RGP. These analyses allow us to gain a deeper understanding of the participants' rationale for or against the adoption of randomly generated passwords. Using the codes reported in Zibaei et al. [51], two researchers in our lab independently categorized users' comments and their rationale. Participants who provided multiple reasons for their password creation behavior were given multiple codes. To ensure the consistency of our coding process, we used Cohen's Kappa to measure inter-rater agreement. The Cohen's Kappa score was $\kappa = 0.95$, demonstrating "almost perfect agreement" between the two coders.

4 Results

Here we describe our participant demographics, discuss the efficacy of Safari's nudge elements and our enhancements, and present factors and reasons behind adoption and abandonment behavior.

4.1 Participant Demographics

Table 9 presents an overview of the demographics of our participants across all conditions. A total number of 896 participants initially signed up for our study, where 34 participants opted out before debriefing. From the remaining 862 participants, 9 respondents were removed due to being duplicates or failing in answering the attention question. Some notable statistics of our study across all conditions (not reported in Table 9) follow. Participants identified as 58.7% male, 40.7% female, and 0.6% preferred not to answer. Participants were mostly 26-35 years old. Of their education, 68.1% had a bachelor's degree and almost half of our participants had a business and IT background.

4.2 Safari's Design or its Users? (RQ1)

Zibaei et al. [51] only evaluated Safari's nudge among Safari users. This motivates us to ask whether Safari's nudge design remains more effective for users of another browser? In particular, among Chrome users, does Safari's nudge design remain more effective than Chrome's? We compare the RGP adoption rates between condition groups Chrome and Safari (67.6% vs 81.1%). Using the χ^2 test, we reject the null hypothesis ($\chi^2 = 7.95$, $p = 0.0038$) with small effect size (Cramer's $V = 0.16$). See Table 3 for more details. This result implies that Chrome users are more likely to adopt a randomly generated password when exposed to Safari's nudges compared to Chrome's nudges, supporting that the effectiveness of Safari is likely due to its design.

4.3 Which Nudge Elements? (RQ2)

Which elements of Safari's nudge contribute to its high RGP adoption rate? To answer this question, we first attempt to understand whether the nudges involved in PwdField are more effective than those of Pop-up (see Figure 1). So, we compare RGP adoption rates between condition groups PwdField and Pop-up (75.2% vs 57.9%). Using the χ^2 test, we reject the null hypothesis ($\chi^2 = 9.33$, $p = 0.0022$) and the effect size is weak (Cramer's $V = 0.18$). Thus, users are more likely to use a randomly generated password when they are exposed to PwdField (which employs default, deceptive visualization, suggesting alternative, and just-in-time prompt

Table 3: χ^2 test results indicate that Safari's nudge is significantly more effective than Chrome's even for Chrome users, and PwdField is significantly more effective than Pop-up. PwdField-No-Visual and PwdField are statistically comparable in their nudging ability. The Safari-Social nudge offered no significant improvements over Safari's nudge. *Significance level $\alpha = 0.005$.

Research Question	Test Description	df	N	χ^2	p	V
RQ1	Chrome vs. Safari	1	282	7.95	0.0038*	0.16
RQ2	PwdField vs. Pop-up	1	279	9.33	0.0022*	0.18
	PwdField vs. Safari	1	296	2.386	0.122	N/A
	Safari vs. Pop-up	1	269	19.657	< 0.00001*	0.27
	Chrome vs. PwdField-No-Visual	1	281	9.077	0.002*	0.17
	Chrome vs. Pop-up	1	265	2.66	0.102	N/A
	PwdField vs. PwdField-No-Visual	1	295	2.79	0.09	N/A
RQ3	Safari-Social vs. Safari	1	293	0.58	0.80	N/A

Table 4: RGP adoption rate across prototypes

	PwdField-No-Visual	**Safari**	**Safari-Social**	**PwdField**	**Chrome**	**Pop-up**
RGP adoption rate	83.1%	81.1%	80.0%	75.2%	67.7%	57.9%

nudges) compared to Pop-up (which employs only suggesting alternative and just-in-time prompt nudges).

One might wonder if the deceptive visualization nudge in visual effect (i.e., Highlight and Fade) increases the RGP adoption rates for PwdField. To test this, we compare the RGP adoption rates between condition groups of PwdField and PwdField-No-Visual (75.9% vs 83.1%), which don't exhibit a statistically significant difference ($\chi^2 = 2.79$, $p = 0.09$). Thus, the Visual element doesn't appear to improve RGP adoption rates over just the Autofill element alone (one can even note the higher adoption rate of 83.1% for PwdField-No-Visual). Table 3 highlights interesting findings obtained through pairwise comparisons between conditions. Taken together, these results indicate that Autofill is the most important element of Safari's nudge. Autofill is the only part of the interface that implements a *default nudge*. It provides evidence that using a simple default is the most effective type of nudge for encouraging RGP adoption. However, other UI elements might have other advantages as discussed in Section 5.

4.4 Can a Social Nudge Improve? (RQ3)

Can we further extend Safari's RGP adoption rate by incorporating a social influence nudge? We compare RGP adoption rates between condition groups Safari and Safari-Social (81.1% vs 80.0%). The null hypothesis holds true ($\chi^2 = 0.58$, $p = 0.8$). Thus, users exhibit a similar likelihood of selecting RGPs when they are exposed to Safari and Safari-Social. We conclude that our prototype, which was a nudge to enable social comparisons, could not enhance the Safari nudge's ability to encourage more users to adopt RGPs.

4.5 Contributing External Factors

Do the external factors identified in other studies [51] (i.e., nudge noticeability and previous experience using RGPs) impact RGP adoption rates in our study? We compare RGP adoption rates of the users who self-identify as having previously used RGPs and who had not (78.8% vs. 20.0%). We reject the null hypothesis ($\chi^2 = 37.44$, $p < 0.0001$) with small effect size (Cramer's $V = 0.20$). This suggests that prior experience with randomly generated passwords influences their adoption and usage.

We also compare RGP adoption rates of the users who answered yes to the same post-study question as other work [51]: "Did you notice the recommendation to use a random password while registering on our website?" and those who reported had not (63.7% vs. 2.87%). We reject the null hypothesis ($\chi^2 = 26.97$, $p < 0.0001$), with small effect size (Cramer's $V = 0.17$). This suggests that participants who noticed the nudge were more likely to adopt a RGP compared to the participants who did not notice the nudge. We examine the issue of noticeability in more detail, for each prototype, in Section 5.

4.6 RGP Adoption Reasons and Barriers

Why do people adopt or reject RGPs? To gain insight, we administered a post-study questionnaire to ask why they either selected or rejected the RGP by asking "Can you describe the reason why you used/did not use the random password generator?". The results (see Tables 5 and 6) revealed that the main reason for adopting an RGP is the security it offers (35.76% of total participants). Convenience is the second most common adoption reason (13.72% of total participants). Interestingly, security concerns are the primary barrier to

RGP adoption (8.91% of total participants). Most security concerns refer to password manager's potential vulnerabilities related to password vault breaches, and privacy and safety issues. The second most common rejection reason is the issue of memorability. Participants preferred selecting memorable passwords to ease their use across multiple devices. However, this perception of being unable to use password managers across multiple devices may stem from a lack of knowledge regarding the functionality of password managers. It may also be due to not using the same browser across multiple devices, or simply not trusting them to sync.

5 Discussion

We discuss the interpretations of our findings, some more exploratory findings, as well as other considerations of interest. In particular, we examine the interplay between RGP adoption and users noticing the nudge in Section 5.2. We discuss the ethics of default nudges and value of other nudge types in Section 5.3. We end this section with a discussion of limitations in Section 5.4.

5.1 The Power of Simple Default Nudges

By dissecting Safari's nudge and examining its element's efficacy in nudging, we find that Autofill is the most powerful at encouraging RGP use. Autofill is implementing a simple default nudge by automatically filling in a suggested RGP for the user. This simple default creates some friction for users who wish to choose their own password, since they need to either navigate to the "Don't Use" button or manually delete the RGP. It also prominently reinforces that keeping the RGP is the recommended action.

A closer examination of the rates of RGP adoption across all conditions also shows that the prototypes incorporating the default nudge (Safari, Safari-Social, PwdField, and PwdField-No-Visual) have RGP adoption rates that are 75-83%, whereas the prototypes that do not use a default nudge (Chrome and Pop-up) have RGP adoption rates between 58-68%. Grouping all data from conditions that incorporate a default nudge vs. those conditions that do not incorporate default nudges, we find the presence of a default nudge is more effective at encouraging RGP adoption ($\chi^2 = 28.27$, $p < 0.0001^*$, $V = 0.18$).

Our work supports that default nudges are quite powerful at encouraging RGP use. Our findings are in line with reviews that found default nudges are one of the most effective types of nudge across different domains and applications [17]. We believe this is good news for deployment of security nudges in general, as default nudges are easy to implement, and have less parameters to adjust in the design that can lead to its success or failure. Even something as simple as a pop-up that intends to suggest alternatives can fail due to subtle choices in words, colors, positioning, etc. Visualizations can be even more challenging to design. However, such elements may improve default nudge designs from an ethical perspective (see Section 5.3 for further discussion).

5.2 Is Noticing the Nudge Important?

We explore whether *noticeability* (i.e., how noticeable a nudge is) might be a cause of some prototypes being more effective than others. For each prototype, Table 7 shows the relationship between noticing the nudge and RGP adoption. While we found in Section 4.5 that participants (across all conditions) who noticed the nudge were more likely to adopt an RGP, a closer inspection using Pearson correlation reveals that RGP adoption is only positively correlated with noticing the nudge in the Chrome and Safari prototypes. All other conditions had no noticeable correlation and in one prototype (PwdField), negative correlation. The positive correlation only being in the Chrome and Safari groups implies that familiarity with the interface can lead to higher trust and subsequent RGP adoption. A closer examination of the other prototype nudges (which are all novel to Chrome users, since they are neither exactly the Chrome or Safari nudge) reveals some interesting insights. In particular, Pop-up was comparably noticeable to the other conditions, yet had a significantly lower rate of adoption; it was the only novel prototype not involving a default nudge. Another interesting comparison point is PwdField vs. PwdField-No-Visual. When the visual element was missing, more participants chose to adopt the RGP (both in the group that noticed the RGP and who did not notice the RGP). One possible reason is the interface drawing less attention to itself and therefore caused fewer participants to hesitate and seriously weigh their options. We discuss this issue further in Section 5.3.

We analyzed user's comments to gain a deeper understanding of why some individuals chose to reject the RGP even after noticing the nudge. The most common barrier among participants who acknowledged the nudge but rejected the RGP was security concerns (22.75% of the participants). The second most commonly mentioned barrier was the difficulty of memorizing the RGP (17.96% of participants). These reasons (and their percentages) are comparable to all users who rejected the RGP, regardless of whether they noticed the nudge, so it appears that noticing the nudge is simply not enough to change some users' beliefs about the security offered by RGPs and usability of password managers.

5.3 Ethics of Default Nudges

We were surprised to observe that PwdField-No-Visual was more effective than PwdField—we had expected that the Visual element of PwdField would make the prototype more visually striking, leading to higher rates of RGP adoption. One possible explanation is that by the interface drawing less attention to itself, fewer participants took enough notice to seriously consider the implications of adopting the RGP.

Table 5: Reasons for Adopting the RGP

Code	N	%	Sample of Comments
Security	305	35.7%	*"I used it because it gave me a strong password"*
Convenience	117	13.7%	*"It was easier than coming up with my own password."*
Noise	100	11.7%	*"Z9tOh\ES*GOX"*
User preference	49	5.7%	*"I always use a random password generator."*
Incongruous	48	5.6%	*"I'd rather make my own"*
Remember password feature	38	4.4%	*"I always do them when I can and save it to my computer/Google for ease of login"*
Didn't care about the website	11	1.3%	*"I thought I should because I am doing a HIT."*
Unsure	8	0.9%	*"I don't know how to use it and have never really heard of it until now."*
Strict password policy	3	0.3%	*"Use random password generator because we can't match the requirement."*

Table 6: Reasons for Rejecting the RGP

Code	N	%	Sample of comments
Security concern	45	5.8%	*"The random password generator wasn't running locally on the CPU; so it was insecure."*
Memorability issue	43	5.0%	*"I would rather use a password that I can memorize.*
Noise	39	4.7%	*"None"*
Incongruous	35	4.1%	*"Using the random password is very difficult to hack"*
Trust issue	30	3.5%	*"I did not trust it"*
User preference	26	3.0%	*"I can create my own password"*
Didn't care about the website	8	0.9%	*"Not sure if I will keep this account."*
Didn't notice the nudge	6	0.7%	*"I did not see that option"*
The desire to reuse password	4	0.4%	*"I like to use similar passwords for each website."*
Lack of knowledge	2	0.2%	*"I was unaware of it."*

This, combined with the observed higher efficacy rates of the default nudges, raises the question of whether default nudges have ethical considerations? What if the user doesn't stop to consider the implications of accepting the default? If the user fails to notice that a default has been set, which they have a choice to accept or reject, then has something unethical occurred (even if it is the "best choice")?

From Table 7, we notice that for most conditions, the percent of users who didn't notice the nudge and rejected the RGP is higher than the percentage of users who noticed the nudge and rejected the RGP. For the PwdField and PwdField-No-Visual groups, this effect is reversed with most of the users who didn't notice the nudge accepting the RGP. These are the two groups that employ default nudges but not pop-up messages. This observation raises the question of whether pop-up messages draw a user's attention that they need to make a decision. Given these observations, Pop-up and Visual elements, while not more effective at encouraging RGP adoption, have advantages from an ethical perspective. We believe further research on nudging should also consider additional metrics of success. For example, in the context of RGP nudges, perhaps to consider the user's understanding of the decision they made and its implications.

5.4 Limitations

We caution readers against interpreting our raw percentages as rates of RGP adoption in other non-experimental settings. This is due to the prevalence of low-quality data from the Amazon MTurk platform [23]. However, the comparisons between the different conditions we test should have validity, as the amount of low-quality data (or noise) should be similar between each of the conditions we test. To reduce the impact of poor data quality, we add a question that aims to catch inattentive participants, which we have excluded from our analysis.

For all non-Chrome conditions (e.g., Safari, PwdField, etc.), Chrome users might have noticed the interface was different than usual. One might ask whether this could explain Safari's higher RGP adoption rate, since this novelty might have brought additional salience to the nudge. To this end, we examine the relationship between noticing the nudge and RGP adoption in Section 5.2. Our findings in that section indicate that Chrome and Safari conditions have similar noticeability ($\chi^2 = 3.59$, $p = 0.057$) and some Chrome users (20.1%) rejected the RGP despite noticing the nudge. Thus, appears that Safari's higher success rate is not because of being more noticeable, but in nudging users towards accepting the RGP after gaining their attention. To further reflect on whether novelty could explain our result, we discuss the success rates of our other prototypes that should be novel to Chrome users.

Table 7: For each prototype, the relationship between noticing the nudge (Noticed Y/N) and adopting the RGP (Y/N). Percentages are shown as well as Pearson correlation values (r).

		Chrome		Safari		PwdField		PwdField-No-Visual		Pop-up		Safari-Social	
RGP		Y	N	Y	N	Y	N	Y	N	Y	N	Y	N
Noticed	Y	63.76%	20.1%	77.6%	13.28%	65.35%	22.78%	77.46%	12.67%	52.38%	34.12%	76.66%	16.66%
	N	2.87%	10.79%	4.19%	4.19%	9.80%	1.96%	5.63%	2.81%	5.55%	7.14%	3.33%	2.66%
Cor. r		0.4		0.24		-0.69		0.14		0.11		0.16	

Table 8: For each prototype, the percent of users who reported: accepting the RGP due to beliefs it was secure (Security), and rejecting the RGP due to trust or security concerns (Mistrust).

	Chrome		Safari		PwdField		PwdField-No-Visual		Pop-up		Safari-Social	
RGP	Y	N	Y	N	Y	N	Y	N	Y	N	Y	N
Security	48.95%	N/A	46.61%	N/A	42.06%	N/A	44.27%	N/A	54.66%	N/A	37.50%	N/A
Mistrust	N/A	32.60%	N/A	13.63%	N/A	12.5%	N/A	25.00%	N/A	13.20%	N/A	44.00%

We observe that the Pop-up prototype, despite being a novel interface, did not result in higher RGP adoption than Chrome (in fact it was nearly 10% lower).

One might wonder if novelty has introduced a lack of trust in the interface. However, the number of participants who cited trust or security concerns as reasons for rejecting the RGP are overall quite low (see Table 8). Notably, the percentage of participants who cited mistrust was higher in Chrome than in the other conditions (except Safari-Social), indicating that this issue was not prevalent among non-Chrome prototypes/conditions. Despite these observations, it remains possible that novelty may have somewhat increased Safari's nudge success. However, our comparisons between the elements of Safari's nudge interface should be equally impacted by novelty.

We implemented a particular social nudge design in our study, which shows no significant effect toward further improving Safari's nudge design (in terms of RGP adoption). Our result does not suggest the inefficiency of social nudges in general, but rather indicates that our specific design did not produce the desired effect.

Our study is limited to the evaluation of password nudges solely within the context of web browsers; it is possible that results may differ between desktop and mobile devices. It is also worth noting that the wording of the messaging in Safari and Chrome's UI has changed since our study was conducted.

Our study is conducted with a limited diversity of participants on the Amazon MTurk platform, where all of our participants were from the United States and are fluent in English, which may have resulted in a language or cultural bias. While it is shown that MTurk workers are more tech-savvy and younger, previous research implies that online privacy and security behavior studies can still estimate the general population's behavior [35]. However, MTurk users may encounter more account creation scenarios than the general population, leading to a higher rate of RGP adoption.

We collect data on users' behavior when signing up once to test the usability of the registration page. However, users' behavior might differ when the user signs up with the intention of long-term use. Further study is needed to determine whether planned long-term use might reduce RGP adoption rates.

Some users may have used other methods to generate random passwords; as such, we record entered passwords. Our analysis reveals that 8.1% of users demonstrated such behavior, with the following breakdown: Safari (2.8%), Chrome (10.8%), Pop-up (7.9%), PwdField (9.2%), PwdField-No-Visual (8.5%), and Safari-Social (9.3%).

6 Conclusion

Our work provides clarity on a number of issues brought up in other research on password manager RGP nudges. In particular, we offer evidence that Safari's password manager nudge is more effective at encouraging RGP adoption than Chrome's. Additionally, we find which nudge types are most effective at encouraging RGP use—it turns out that simple default nudges are the most powerful. While the other nudges we studied were less effective at encouraging RGP use, they may still serve an ethical purpose in increasing awareness to users regarding their decision.

Future work includes addressing the "missed opportunity" observed in many of the nudge prototypes we studied, where the nudge was noticed but unfortunately failed to capitalize on the user's attention. Future attempts at improving these nudge designs should focus on addressing the main barriers

we identified: concerns about security/trust and the possibility of needing to remember the RGP. Educating users about how password managers work might help. Future work also includes developing approaches to personalize these security nudges to improve their efficacy.

Acknowledgments

This research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC). We thank Nicholas Hughes for his assistance with our study.

References

[1] Autofill your user name and password in Safari on mac. https://support.apple.com/en-ca/guide/safari/ibrwf71ba236/mac. Accessed: 2023-02-10.

[2] How to generate a secure password in Firefox. https://support.mozilla.org/en-US/kb/how-generate-secure-password-firefox. Accessed: 2023-02-10.

[3] Let Chrome create and save a strong password for your online accounts. https://support.google.com/chrome/answer/7570435?hl=en&co=GENIE.Platform%3DDesktop. Accessed: 2023-02-10.

[4] Market share held by leading desktop internet browsers in the Uunited States from January 2015 to August 2022. https://www.statista.com/statistics/272697/market-share-desktop-internet-browser-usa/. Accessed: 2023-02-15.

[5] Nora Alkaldi and Karen Renaud. Why do people adopt, or reject, smartphone password managers? In *European Workshop on Usable Security*, 2016.

[6] Nora Alkaldi and Karen Renaud. Encouraging password manager adoption by meeting adopter self-determination needs. In *Proceedings of the 52nd Hawaii International Conference on System Sciences*, 2019.

[7] Fahad Alodhyani, George Theodorakopoulos, and Philipp Reinecke. Password managers—it's all about trust and transparency. *Future Internet*, 12(11):189, 2020.

[8] Sal Aurigemma, Thomas Mattson, and Lori Leonard. So much promise, so little use: What is stopping home end-users from using password manager applications? In *Hawaii International Conference on System Sciences*, 2017.

[9] Jannatul Bake Billa, Anika Nawar, Md Maruf Hasan Shakil, and Amit Kumar Das. Passman: A new approach of password generation and management without storing. In *IEEE International Conference on Smart Computing & Communications (ICSCC)*, pages 1–5, 2019.

[10] Ana Caraban, Evangelos Karapanos, Daniel Gonçalves, and Pedro Campos. 23 ways to nudge: A review of technology-mediated nudging in human-computer interaction. In *ACM Conference on Human Factors in Computing Systems (CHI)*, pages 1–15, 2019.

[11] Sonia Chiasson, Paul C van Oorschot, and Robert Biddle. A usability study and critique of two password managers. In *USENIX Security*, 2006.

[12] Robert B Cialdini and N Garde. Influence (vol. 3). *Port Harcourt: A. Michel*, 1987.

[13] Molly Cooper, Yair Levy, Ling Wang, and Laurie Dringus. Subject matter experts' feedback on a prototype development of an audio, visual, and haptic phishing email alert system. *Online Journal of Applied Knowledge Management*, 8(2):107–121, 2020.

[14] Serge Egelman, Andreas Sotirakopoulos, Ildar Muslukhov, Konstantin Beznosov, and Cormac Herley. Does my password go up to eleven? the impact of password meters on password selection. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2379–2388, 2013.

[15] Michael Fagan, Yusuf Albayram, Mohammad Maifi Hasan Khan, and Ross Buck. An investigation into users' considerations towards using password managers. *Human-centric Computing and Information Sciences*, 7(1):1–20, 2017.

[16] Maximilian Golla, Björn Hahn, Karsten Meyer zu Selhausen, Henry Hosseini, and Markus Dürmuth. Bars, badges, and high scores: On the impact of password strength visualizations.

[17] Dennis Hummel and Alexander Maedche. How effective is nudging? a quantitative review on the effect sizes and limits of empirical nudging studies. *Journal of Behavioral and Experimental Economics*, 80, 2019.

[18] Moritz Ingendahl, Dennis Hummel, Alexander Maedche, and Tobias Vogel. Who can be nudged? examining nudging effectiveness in the context of need for cognition and need for uniqueness. *Journal of Consumer Behaviour*, 20(2):324–336, 2021.

[19] Jon M Jachimowicz, Shannon Duncan, Elke U Weber, and Eric J Johnson. When and why defaults influence decisions: A meta-analysis of default effects. *Behavioural Public Policy*, 3(2):159–186, 2019.

[20] David Jaeger, Chris Pelchen, Hendrick Graupner, Feng Cheng, and Christoph Meinel. Analysis of publicly leaked credentials and the long story of password (re-) use. *Hasso Plattner Institute, Universidad de Potsdam. Disponible en https://bit. ly/2E7ZT01*, 2016.

[21] Daniel Kahneman, Stewart Paul Slovic, Paul Slovic, and Amos Tversky. *Judgment under uncertainty: Heuristics and biases*. Cambridge university press, 1982.

[22] Ambarish Karole, Nitesh Saxena, and Nicolas Christin. A comparative usability evaluation of traditional password managers. In *International Conference on Information Security and Cryptology*, pages 233–251. Springer, 2010.

[23] Ryan Kennedy, Scott Clifford, Tyler Burleigh, Philip D Waggoner, Ryan Jewell, and Nicholas JG Winter. The shape of and solutions to the mturk quality crisis. *Political Science Research and Methods*, 8(4):614–629, 2020.

[24] Shelia M Kennison and D Eric Chan-Tin. Predicting the adoption of password managers: A tale of two samples. *TMS Proceedings 2021*, 2021.

[25] Shelia M Kennison, Ian T Jones, Victoria H Spooner, and D Eric Chan-Tin. Who creates strong passwords when nudging fails. *Computers in Human Behavior Reports*, 4:100132, 2021.

[26] Thomas C. Leonard, Richard H. Thaler, and Cass R. Sunstein. Nudge: Improving decisions about health, wealth, and happiness, 2008.

[27] Michael D Leonhard and VN Venkatakrishnan. A comparative study of three random password generators. In *IEEE International Conference on Electro/Information Technology*, pages 227–232, 2007.

[28] Raymond Maclean and Jacques Ophoff. Determining key factors that lead to the adoption of password managers. In *IEEE International Conference on Intelligent and Innovative Computing Applications (ICONIC)*, pages 1–7, 2018.

[29] Peter Mayer, Collins W Munyendo, Michelle L Mazurek, and Adam J Aviv. Why users (don't) use password managers at a large educational institution. In *USENIX SOUPS*, 2022.

[30] Sean Oesch and Scott Ruoti. That was then, this is now: A security evaluation of password generation, storage, and autofill in browser-based password managers. In *USENIX Security Symposium*, 2020.

[31] Zach Parish, Amirali Salehi-Abari, and Julie Thorpe. A study on priming methods for graphical passwords. *Journal of Information Security and Applications*, 62:102913, 2021.

[32] Sarah Pearman, Shikun Aerin Zhang, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. Why people (don't) use password managers effectively. In *USENIX SOUPS*, 2019.

[33] Hirak Ray, Flynn Wolf, Ravi Kuber, and Adam J Aviv. Why older adults (don't) use password managers. In *USENIX SOUPS*, 2021.

[34] Elissa M Redmiles, Yasemin Acar, Sascha Fahl, and Michelle L Mazurek. A summary of survey methodology best practices for security and privacy researchers. Technical report, 2017.

[35] Elissa M. Redmiles, Sean Kross, and Michelle L. Mazurek. How well do my results generalize? comparing security and privacy survey results from mturk, web, and telephone samples. In *IEEE Symposium on Security and Privacy (SP)*, pages 1326–1343, 2019.

[36] Karen Renaud, Verena Zimmerman, Joseph Maguire, and Steve Draper. Lessons learned from evaluating eight password nudges in the wild. In *The LASER Workshop: Learning from Authoritative Security Experiment Results (LASER 2017)*, pages 25–37, 2017.

[37] Karen Renaud and Verena Zimmermann. Nudging folks towards stronger password choices: providing certainty is the key. *Behavioural Public Policy*, 3(2):228–258, 2018.

[38] Karen Renaud and Verena Zimmermann. Encouraging password manager use. *Network Security*, 2019(6):20–20, 2019.

[39] Sunyoung Seiler-Hwang, Patricia Arias-Cabarcos, Andres Marin, Florina Almenares, Daniel Diaz-Sanchez, and Christian Becker. I don't see why I would ever want to use it, analyzing the usability of popular smartphone password managers. In *ACM Conference on Computer and Communications Security (CCS)*, pages 1937–1953, 2019.

[40] James Simmons, Oumar Diallo, Sean Oesch, and Scott Ruoti. Systematization of password manager use cases and design paradigms. In *Annual Computer Security Applications Conference*, pages 528–540, 2021.

[41] Elizabeth Stobert and Robert Biddle. A password manager that doesn't remember passwords. In *New Security Paradigms Workshop*, pages 39–52, 2014.

[42] Elizabeth Stobert, Tina Safaie, Heather Molyneaux, Mohammad Mannan, and Amr Youssef. Bypass: Reconsidering the usability of password managers. In *International Conference on Security and Privacy in Communication Systems*, pages 446–466. Springer, 2020.

[43] Kurt Thomas, Jennifer Pullman, Kevin Yeo, Ananth Raghunathan, Patrick Gage Kelley, Luca Invernizzi, Borbala Benko, Tadek Pietraszek, Sarvar Patel, Dan Boneh, et al. Protecting accounts from credential stuffing with password breach alerting. In *28th USENIX Security Symposium*, pages 1556–1571, 2019.

[44] Julie Thorpe, Muath Al-Badawi, Brent MacRae, and Amirali Salehi-Abari. The presentation effect on graphical passwords. In *ACM Conference on Human Factors in Computing Systems (CHI)*, pages 2947–2950, 2014.

[45] Anthony Vance, David Eargle, Kirk Ouimet, and Detmar Straub. Enhancing password security through interactive fear appeals: A web-based field experiment. In *IEEE International Conference on System Sciences*, pages 2988–2997, 2013.

[46] Tina AG Venema, Floor M Kroese, Jeroen S Benjamins, and Denise TD De Ridder. When in doubt, follow the crowd? responsiveness to social proof nudges in the absence of clear preferences. *Frontiers in psychology*, 11:1385, 2020.

[47] Shengqian Wang, Amirali Salehi-Abari, and Julie Thorpe. Pixi: Password inspiration by exploring information. *arXiv preprint arXiv:2304.10728*, 2023.

[48] Yang Wang, Pedro Leon, Kevin Scott, Xiaoxuan Chen, Alessandro Acquisti, and Lorrie Cranor. Privacy nudges for social media: an exploratory facebook study. In *International Conference on World Wide Web*, 2013.

[49] Jeff Yan and Dearbhla McCabe. Gender bias in password managers. *arXiv e-prints*, pages arXiv–2206, 2022.

[50] Iryna Yevseyeva, Charles Morisset, and Aad van Moorsel. Modeling and analysis of influence power for information security decisions. *Performance Evaluation*, 98:36–51, 2016.

[51] Samira Zibaei, Dinah Rinoa Malapaya, Benjamin Mercier, Amirali Salehi-Abari, and Julie Thorpe. Do password managers nudge secure (random) passwords? In *USENIX SOUPS*, 2022.

[52] Verena Zimmermann and Karen Renaud. The nudge puzzle: matching nudge interventions to cybersecurity decisions. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 28(1):1–45, 2021.

Appendix A Demographic Information

Table 9: Demographic information across all conditions.

		Condition					
		Chrome	**Safari**	**PwdField**	**PwdField-No-Visual**	**Pop-up**	**Safari-Social**
		n=139	n=143	n=153	n=142	n=126	n=150
Gender	Female	28.1%	40.6%	45.1%	46.5%	38.1%	44.9%
	Male	71.2%	58.7%	54.2%	53.5%	61.1%	54.4%
	N/A	0.7%	0.7%	0.7%	0.0%	0.8%	0.7%
Age	18-25	19.4%	15.4%	20.9%	24.6%	24.7%	18.8%
	26-35	63.3%	49.6%	45.1%	42.3%	46.8%	34.1%
	36-50	14.4%	23.8%	22.8%	26.8%	21.4%	31.9%
	50+	2.2%	10.5%	10.5%	6.3%	7.1%	14.5%
	N/A	0.7%	0.7%	0.7%	0.0%	0.0%	0.7%
Education	High school	5.1%	9.1%	11.8%	4.2%	8.7%	10.9%
	Bachelor's	71.9%	67.8%	64.1%	75.4%	61.1%	70.3%
	Master's	19.4%	21.0%	22.1%	19.7%	30.2%	17.4%
	PhD/higher	2.2%	1.4%	1.3%	0.0%	0.0%	0.7%
	N/A	1.4%	0.7%	0.7%	0.7%	0.0%	0.7%
Study/Work	Social Sci.& Humanities	8.7%	4.2%	5.2%	8.5%	5.6%	10.2%
	Science	2.2%	5.6%	1.3%	0.7%	4.0%	5.3%
	Health Science	8.0%	16.1%	19.0%	12.0%	17.5%	12.4%
	Engineering & Applied Sci.	12.3%	7.0%	9.8%	7.7%	7.1%	4.3%
	Energy & Nuclear Sci.	1.4%	1.4%	2.0%	1.4%	1.6%	2.2%
	Education	9.4%	10.4%	6.5%	6.3%	7.8%	9.4%
	Business & IT	53.8%	49.7%	49.0%	57.0%	50.0%	47.8%
	Other	2.8%	3.5%	6.5%	6.4%	5.6%	7.0%
	N/A	1.4%	2.1%	0.7%	0.0%	0.8%	1.4%
Language	English	100.0%	97.2%	99.3%	100%	98.4%	97.1%
	Other	0.0%	2.1%	0.0%	0.0%	1.6%	2.2%
	N/A	0.0%	0.7%	0.7%	0.0%	0.0%	0.7%

Adventures in Recovery Land: Testing the Account Recovery of Popular Websites When the Second Factor is Lost

Eva Gerlitz
Fraunhofer FKIE

Maximilian Häring
University of Bonn

Charlotte Theresa Mädler
University of Bonn

Matthew Smith
University of Bonn, Fraunhofer FKIE

Christian Tiefenau
University of Bonn

Abstract

Literature on two-factor authentication (2FA) lists users' fear of losing the second factor as one major constraint on acceptability. Nonetheless, more and more services offer or even enforce 2FA. Yet, little is published about what services do to prevent users from losing access to their accounts and how well users are guided through the process of regaining access to their accounts in case they lose their second factor. To fill this gap, we set up 2FA on 78 popular online services and apps and analyzed their user interface during the 2FA setup and recovery. Although there is no straightforward solution for account recovery when using a second factor, we identified easily fixable usability flaws. For example, in the setup phase, 28 services do not mention the possibility of losing the second factor at all. Furthermore, while it is common for services to provide a clearly visible "forgotten password"-link beneath the login field, an equivalent for 2FA is often missing, and a user is left alone with the problem. Our study provides insights for website designers and security practitioners seeking to enhance the usability of 2FA. We also discuss further directions for research.

1 Introduction

Two-factor authentication (2FA) is one powerful solution to improve account security. In 2FA, a second factor (*secondary authenticator*) is needed to confirm the user's identity. Typically, this second factor is something the user *is* or *has* [21]. This is used in addition to the *primary authenticator*, typically something the user *knows*.

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.
USENIX Symposium on Usable Privacy and Security (SOUPS) 2023.
August 6–8, 2023, Anaheim, CA, USA

Using such a second factor is one of the most frequently given advice experts give non-tech-savvy users to stay safe online [5, 24], and indeed, the use of 2FA rose steadily over the last years [7]. Some services even force users to secure their accounts with second factors [19] or are required by law to do so, e.g., banking websites in the EU [34].

To understand the consequences of this additional security mechanism from the users' perspective, several studies examined the usability of (possible) second factors (e.g., [1,8,30,38,39]), their initial setup (e.g., [2,10,39]) , or looked at the acceptability of 2FA (e.g., [9, 10, 39, 43]).

Within these studies, participants repeatedly expressed the fear of losing the second factor [10, 25, 35] and statistics indicate that around 40% of smartphone users have had at least one incident in which they lost their device or had it stolen [3, 23, 27]. Considering that the personal smartphone is a convenient choice for 2FA [41], these numbers indicate that many users might find themselves in a situation where they no longer have access to their second factor and therefore be locked out of their account. The consideration of being locked out of a personal account can lead to a low acceptance of 2FA [10]. However, little work has been conducted to understand how services deal with the threat of their users being locked out.

In this work, we want to understand how websites and apps, as one major use case for 2FA, guide a user through the *setup* of 2FA and the *recovery* after losing the second factor. Specifically, we were guided by the following research questions:

RQ1: (How) do popular services communicate the issue of losing the second factor to their users? I.e., do they communicate the issue? Do services encourage users to set up another factor as a backup? Do they provide backup codes? Is the user forced to do something, e.g., downloading backup codes?

RQ2: How well are users supported through the services' recovery protocol when they try to log in but the second factor is lost? I.e., do users receive help during login if their second factor is not accessible anymore? What are

their options?

RQ3: What information do users need to provide to regain access to accounts? I.e., is personal identification needed? Does the user need to have information about the account's activities?

To answer the research questions, we conducted 78 expert reviews that focused on the current practice of online services. We created accounts, enabled 2FA, and analyzed the services' way of informing the user about the possible risks of enabling 2FA and what a user can do to mitigate them. We then ran through the account recovery processes without the second factor and without backup codes. We captured how the service led through this process, what was needed to recover the account, and whether recovery was possible at all.

Overall, we were able to gain access to half of the accounts. This low number might be well explained by security reasons but indicates that users' naive assumptions when they lose their second factor should not be that they could regain access as easily as they would in the case of a forgotten password.

Our results show that the investigated services do not share a common practice, neither during 2FA setup nor during recovery. Looking at the setup, 20.5% of the services do not seem to provide any backup possibilities at all; on the other hand, 20.5% of the services force the user to implement a fallback for the second factor or download backup codes. Only 12.8% of the services clearly communicate that the user will lose access to the account without the second factor or access to fallback authentication.

The same heterogeneity applies to the process of *recovery*: 19.2% of the services offer the user to use backup codes or alternative ways to receive the needed code during login and additionally link to a direct contact possibility if backups do not work either. On the other side of the spectrum, 17.9% of the services do not help the user at all during login, and the only possibility a user has is to cancel their login attempt and try to find a solution on their own (e.g., by looking at the website's FAQs).

Several of the issues we identified can easily be fixed. We suggest establishing a more standardized approach to 2FA setup and recovery to ensure convenience for their users without impacting security.

2 Related Work

This section summarizes work relevant to our study. We first look at the motivation of our work and the frequency users lose a second factor, followed by studies that analyzed the protocols of different aspects of account recovery on websites.

2.1 Losing Access by Losing a Second Factor

The fear of losing a device that is needed to log in, e.g., as a second factor, and losing access to the account, in general, is mentioned as one major constraint on the acceptability of 2FA in several studies (e.g., [10, 13, 25, 35]). Sometimes, this is accompanied by the fear of impersonation attacks after the loss or theft of this device [35]. Despite this fear, the results of a study by Das et al. [10] indicate that websites might not communicate the issue of loss well during the setup process: Participants were requested to add a security key to their email accounts and were explicitly asked what they would do if they lost the key afterward. Almost a fourth of the participants did not know how to recover this newly set up Yubikey in case it got lost or stolen. Yet, we are unaware of any study investigating how websites communicate a potential loss during login and whether users are nudged or forced to set up another factor as a backup login possibility. We fill this gap with RQ1. Additionally, we want to understand how justified this repeatedly mentioned fear of consequences of losing the second factor is by testing how easy a user could regain access to their account (RQ2 & RQ3).

How Likely is it to Lose the Second Factor? In the following, we report on how often users are confronted with the problem of losing their second factor. This motivates our task design, as we assume that the loss of the second factor is not a theoretical scenario. For smartphones, which are the most commonly used second factor [41], studies indicate that around 40% of smartphone users have had at least one incident in which they lost their device or had it stolen (around 10-15%) [3, 23, 27]. One study estimated that an average person living in the UK loses two smartphones within their lifetime [6]. However, the authors did not report the frequency of users being able to recover their devices: Data from 2014 show that while 90% of phone theft victims tried to recover their phone, only 32% were successful [27]. Furthermore, one study indicates that around 60% of the users who lost their device misplaced it, most often at home or work (49.5%) [22], where chances of finding the device again are high.

Dutson et al. [12], and Abbott et al. [1] looked at implications for the users after their universities adopted 2FA. In the study by Dutson et al. [12], around a fourth of the participants reported they have had at least one incident within one year in which they could not access their account due to an inability to access their phone (because it was lost or stolen, they forgot it somewhere or it ran out of battery). Around 16% of the support chats that were analyzed by Abbott et al. [1] concerned how to access the account if the second factor was inaccessible. For both studies, it remains unclear in how many cases this status was only temporary (i.e., how many people actually lost their device or had it stolen).

So, while we do not have much evidence, we think it is fair to assume that the loss of the second factor, i.e., the smartphone, is something that indeed happens.

2.2 Analysis of Recovery Protocols

We are aware of only a few studies that analyzed the recovery protocols users had to follow if the primary or secondary authenticator was lost or compromised:

Li et al. [26] investigated the recovery protocols for the **primary authentication** for over 200 websites in 2018. They found that on 89.1% of the websites, it was sufficient to have access to the registered email to recover the account. On 4.6%, it was sufficient to know the answer to a security question.

Neil et al. [31] analyzed 57 American websites in 2020 according to their user-facing advice on restoring the user account to a pre-compromise state. For the phase of account recovery, i.e., regaining access to the account **independent of the authentication methods** in use, the authors found that 96% of the websites had some information on what to do (e.g., advising to send oneself a password reset email). Over 60% of the websites recommend contacting their support. Markert et al. [28] extended the previous study by investigating 158 websites; covering the 50 most popular websites in 30 countries. Even though less than in the US American sample, most websites offered some advice on how to recover accounts; mostly by recommending to reset the password or by contacting the support.

Another related study was conducted by Quermann et al. [37], who analyzed the state of user authentication in 2017 for 48 different services (websites, IoT, and mobile devices). They found that none of them offered an easy way to recover accounts that were secured with a **second factor**, and almost all services require the user to contact the services' support.

However, Quermann et al. [37] did not further systematically investigate whether websites do anything to prevent user lockout when users set up a second factor or how well users who cannot access their second factor are guided through the support (e.g., do users have a direct and easy way to contact the support or do they have to search for a contact themselves within various articles?) We update and expand upon this prior work by conducting expert reviews mimicking a user who lost their second factor and analyzing the steps that needed to be taken to regain access, as well as the usability of the support offered by each website/service (RQ2).

3 Methodology

We analyzed how popular services communicate and handle the issue of second-factor loss during the setup and recovery. We did this by conducting 78 expert reviews. The tasks were first to set up a user account with 2FA and, second, to recover it without the factor. In this section, we describe how we selected the evaluated services, the tasks we performed, and how we analyzed the gathered data.

3.1 Service Selection

We used Tranco [36] to identify high-traffic websites and used the top 500 for our analysis. The list was generated on 2 August 2022 [42]. The websites were accessed between September 2022 and January 2023 from Germany with a Linux machine using Chrome. All services that required an app-based setup were accessed from a smartphone (Honor 8x) with Android 8.1.0. During the reviews, we visited the websites as they were referenced on the list. However, in some cases, the websites forwarded us to the localized site according to our location.

We excluded sites if they were marked insecure by Google Safe Browsing or if account creation was only possible for a specific user group. The whole list of exclusion criteria is given in Appendix A.1. An overview of this elimination process and the corresponding numbers is shown in Figure 1.

Websites that belong to the same domain or use shared accounts were merged (e.g., Google.com and YouTube.com). Finally, we checked whether we could enable 2FA on each of those websites. Similar to the findings of Gavazzi et al. [15], less than half of the websites offer 2FA. Finally, we ended up with 78 services for the reviews.

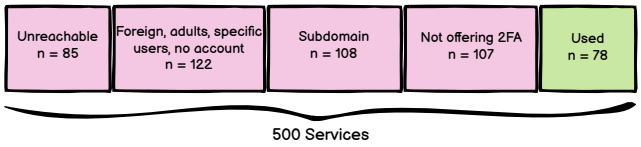


Figure 1: Overview of the service selection. We started with 500 high-traffic websites, according to Tranco [42]. Services were excluded based on criteria specified in Appendix A.1. Eighty-five were unreachable or marked insecure by Google Safe Browsing. This left us with 185 services, of which 78 offered to add a second factor.

3.2 Task

The expert reviews consisted of two tasks. The first task was to create an account and set up a second factor. In the second task, we tried to recover the account, pretending to have no access to the second factor. In the following paragraphs, we describe the tasks in more detail and explain how we conducted the reviews.

Task 1: Setup One researcher manually created accounts on all of the selected services. They always selected the free version of an account and used the same password. They enabled a second factor if possible. For this, they picked the first option allowed based on the following order 1) SMS

verification, 2) email verification,[1] and 3) an authenticator app. If an authenticator app was necessary, they used Google authenticator [44]. The researcher did not set up any additional second factor or possibility to be contacted during the setup phase, except when it was mandatory. The sessions were screen recorded.

After setting up all accounts, the browser was un- and reinstalled to remove artifacts from the setup phase. While this might make it harder to regain access, we opted for the lower-bound results. We believe that if recovery is possible in our scenario, it will also be possible when the browser was already used to log into the account, but not vice versa.

Task 2: Recovery One month after the second factor was added, the same researcher navigated to the login screen and tried to log in without the second factor, i.e., looking for an alternative or help. They did not have access to the backup codes if the service provided them. However, they could answer basic questions about themselves and the account. If 2FA was set up using a smartphone (SMS or authenticator app), the researcher could access the email associated with the account.

If the website gave instructions to regain access, they were followed. If the website did not provide assistance during the login process, the researcher searched through the help center, if any existed, and followed the steps, if any were given. If this also did not help to regain access to the account, the researcher consulted Google with the search term "2fa lost site:www.example.com." If they had to contact support, they used the following text (if applicable): "Hello, I lost my phone, which I use for two-factor authentication, and now I cannot log in. Would it be possible for you to deactivate this, or will I need a new account? Kind regards, [Name]." The recovery was declared successful if it was possible to log in without the second factor, and the second factor could be deactivated or changed. An account was marked as irretrievable if no information could be found on retrieving it, if instructions were given but failed, or if the instructions clearly stated that retrieval was impossible.

The sessions were again screen recorded, and related emails were saved.

3.3 Analysis

To find common themes during the setup and recovery phase, two researchers looked at a random subset of the services (14 services, 18% of all) to create an initial code book for each research question. In this step, each website was represented by all videos and emails associated with the setup and recovery procedure on this particular service (see Section 3.2).

[1] Even though receiving codes through email is not considered as a second factor by NIST [33], it was listed as such on these websites. We opted to go with the definition of the services, as we believe there are users who will do so as well.

The researchers then coded another eleven services (14% of all) using the code book, arriving at a weighted inter-coder reliability of 0.89 which was in the range of 0.56 to 1 for individual codes. For the full coding, each researcher coded half of the services.

4 Results

This section presents the results of the usability evaluation of the 2FA setup and recovery process of 78 services. We first give a general overview of what second factors were supported and recommended by the services. Following this, we show how services try to prevent issues that result from a user losing their second factor during the setup phase (RQ1), e.g., by recommending implementing alternative login methods as backups. In Section 4.3, we report how (well) services guided us through the process of regaining access (RQ2) and what information was needed (RQ3).

A complete overview of all services, the used second factors, and characteristics during setup and recovery are given in Table 2 in Appendix A.

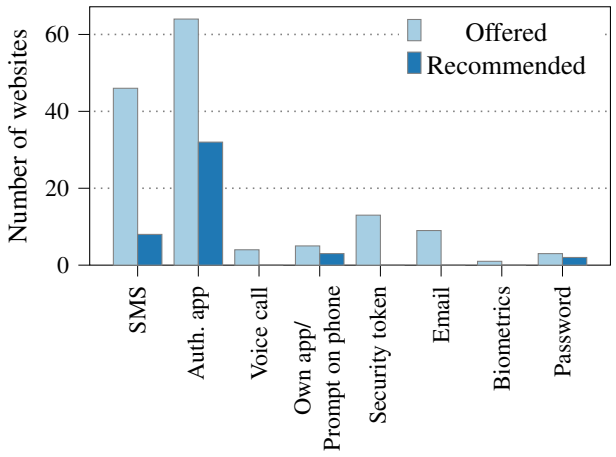


Figure 2: Overview of the allowed second factors on all sampled services. Most services allow users to use an authenticator app. Marked as "recommended" are those factors that were offered as the only possibility, were selected by default, or were marked as "recommended".

4.1 Allowed Second Factors

We were able to add a second factor to 78 services (see Figure 1). We registered a phone number to receive SMS codes on 46 services. If a service did not offer 2FA via SMS, we selected to receive codes via email ($n = 5$) or Google Authenticator ($n = 25$). There was no website where this was not sufficient. In the particular case of two apps where the phone number was already used as a primary authenticator, we added a password as a second factor.

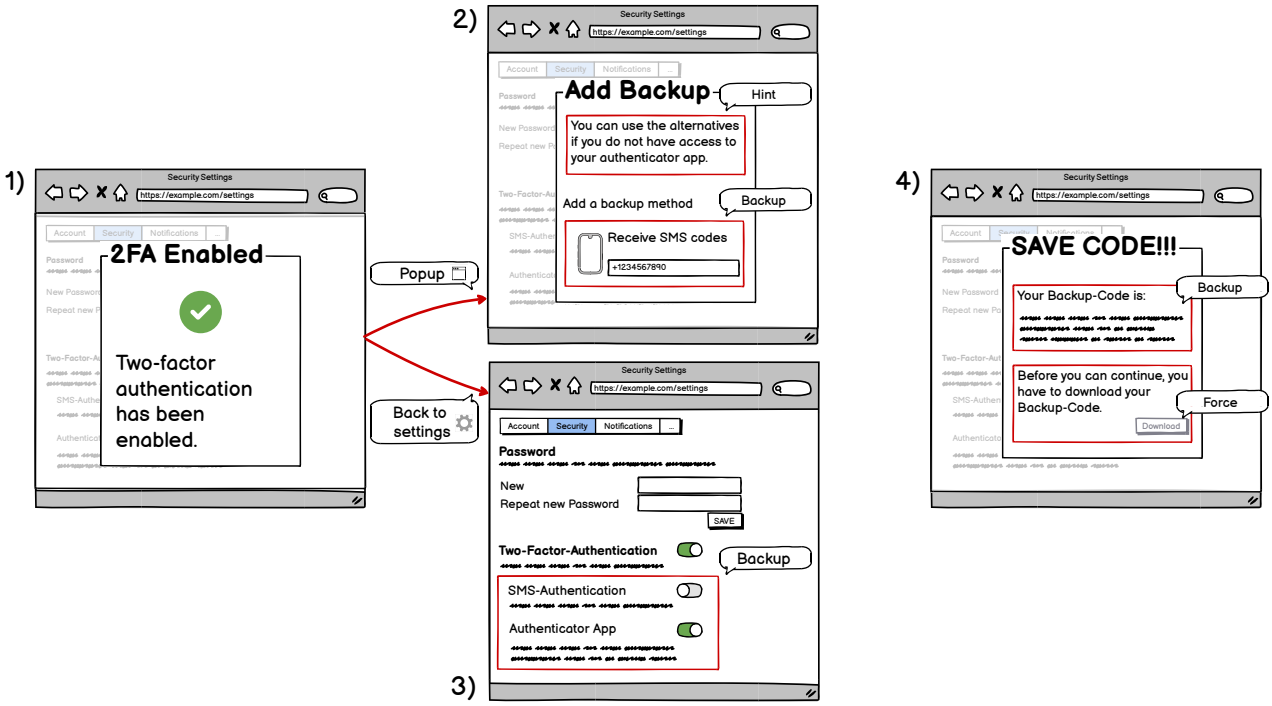


Figure 3: This figure shows four windows, with examples of information and cues we received during the reviews of the setup. A common workflow led us to one of two different states after enabling 2FA (1). In 52 cases (2), hints and backup possibilities were shown in the same popup that was used for setup. Some pages closed the window and led us back to the settings ($n = 10$) (3), where hints and backup possibilities were shown. 16 services required additional action from the users, e.g., requiring them to download backup codes or to add an additional phone number (4).

As shown in Figure 2, most of the investigated services offered the possibility to use authenticator apps to secure user accounts. Authenticator apps were also the most commonly recommended second factor (by 41.0% of the services). Some services mentioned specific authenticator apps, most prominently Google Authenticator ($n = 27$), followed by Authy ($n = 14$) and Microsoft authenticator ($n = 10$).

4.2 2FA Setup

In this section, we report whether and how the services communicated the issue of losing the second factor (RQ1).

For this, we analyzed how prominent they mentioned a potential second-factor loss. We tracked whether and how the services nudged or forced users to add another factor as a backup or store backup codes. The data for this section was gathered during and right after a second factor was added to an account, thus at a point in our scenario where the user still had access to the second factor.

During the analysis, we identified three cues (see Figure 3 for examples) of how services communicate with users related to the research question:

(a) **Hint**: The service mentions that the second factor could be inaccessible.

(b) **Backup**: The service presents possible backup possibilities - backup codes, a security question, or other available factors.

(c) **Force**: The service forces the user to add a backup or download backup codes.

The three cues were shown at one of two locations: Either in the settings ($n = 10$) after the setup of the second factor is completed or in a separate window during or following the setup ($n = 52$).

Backup On most services (79.5%), it was possible to add another alternative second factor ($n = 40$ services) and/or to download one or several backup codes ($n = 45$). Yet, the intended usage of the latter differed: While most services provided backup codes that can be used instead of a code sent by SMS or generated by an app, some services offered a backup code that will automatically deactivate 2FA once used. We found that the wording of these codes differed as well: Both terms “backup codes” and “recovery codes” were used interchangeably, sometimes meaning different things.

Hints Most services that offered backup possibilities (80.6% of the 62 services that offered a backup) hinted at

the possible inaccessibility of the second factor somehow. A typical text was similar to the following: "This code lets you log in if you don't have access to your two-factor authentication methods." In these cases, a user may understand additional factors as a possibility rather than a necessity. Only three websites communicated this a bit more clearly by using statements similar to "you will need these codes should you not have access to your phone." In general, the consequences of loss (i.e., being locked out of the account if losing the second factor and having no access to any backups) were only communicated by a minority: Four services used phrasing similar to: "otherwise you may get permanently locked out." Only ten services clearly stated that the provided backup codes or offered fallback authentication are the "only" way to log in if the second factor is not accessible. Interestingly, this turned out not to be the case for six of these services. We pick this topic up in Section 4.3.2.

Force The use of force was not that common. We only had to add a backup on 16 services. All except one page forcing the user to add a backup explained that this backup could be used to access the account.

Combinations The most common combination of the three cues was to have a hint and backup possibilities but no force to implement them ($n = 29$, 37.2%). The second most common combination was to show and mention nothing at all ($n = 16$, 20.5%): No hint as to what could happen and no way to resolve this. All combinations of the cues are shown in Figure 4. 64.1% of the websites gave a hint and offered a backup possibility.

4.2.1 Tales From the 2FA-Setup Land

We found an interesting case where one page advertised 2FA right after login and also included a small note that one should "remember to create backup verification methods." However, after registering an authenticator app, this information was not shown anymore, though one of the presented verification methods was called "Recovery codes." In another case, the website seemed to follow a more serious approach. After telling the users in the first step to "save this [backup] key," they were told in the second step "Seriously, save this key."

4.2.2 Summary of the Setup Task (RQ1)

To summarize, we were successful in activating 2FA on 78 services. Of these, 50 provided at least minimal information about what to do when the second factor is lost (Hint). Most services offer some form of backup method, and 45 provided backup codes. The degree of how straightforward consequences of loss were communicated differed. Only ten services clearly indicated that a user will lose access to the account without the second factor and without backups. Having all sorts of combinations of hints, backup possibilities, and obviousness, there does not seem to be a process or possibility for fallback authentication a user can assume by default or always rely on.

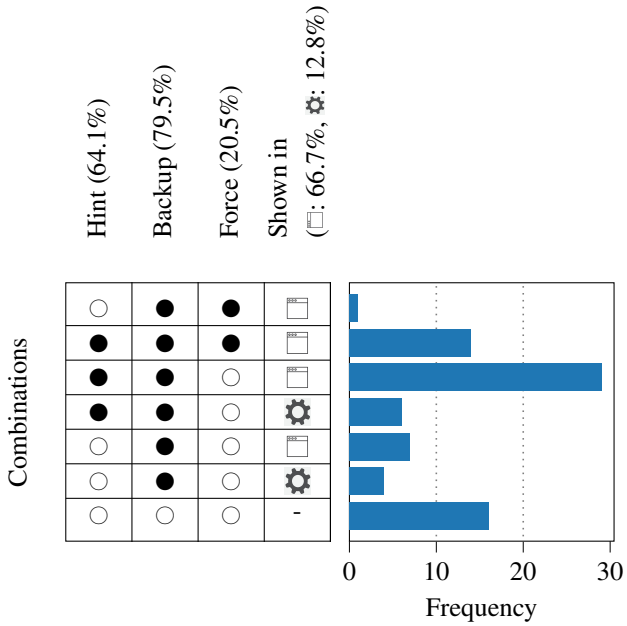


Figure 4: Number of websites that mention the possibility that the second factor is not accessible (Hint) in combination with showing alternative possibilities (Backup) or forcing the user to do something (Force) during 2FA setup. ○: The service does not include the characteristic. ●: The service fulfills the characteristic. "Shown in" depicts the location where this information is shown. □: The information is shown in a popup that also directed us through the process of adding the second factor. ⚙: The information was shown in the settings. Examples are given in Figure 3.

4.3 Recovery

In this section, we present the results for the second task, the account's recovery after the second factor is lost (RQ2).

We looked at how and to what extent the services' interface assisted the user during login, what needed to be done to regain access (e.g., what information had to be provided), and report on how many services we received full access to.

4.3.1 Assistance During Login

We found varying degrees of assistance from the services to guide the user during a login attempt. In the next paragraphs, we clustered common themes.

Missing Common Practice Today, it is common for websites to provide a "forgotten password"-link during login that

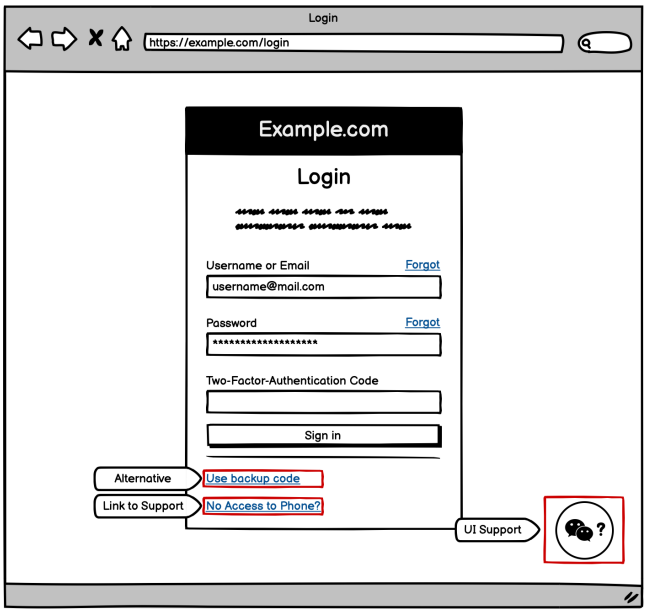


Figure 5: Example for the Login screen. If a link to the support existed, we noted whether it linked to a general FAQ, a specific FAQ (that at least partially mentioned what to do if losing the second factor), or whether a user is provided with an email address or can fill a form.

a user can use to reset their password. As expected, all websites in our set provided such a link. The equivalent for 2FA, i.e., a link a user can click while trying to log in but having no access to the second factor, is often missing. Even though 75.6% of the websites provided the user with some form of a button to offer help in such cases, the usefulness varied massively. Some services mentioned fallback authentication (e.g., suggesting to use backup codes), some linked to some sort of support, and yet others had an always visible support interface that was independent of the login screen. Examples of these possibilities are shown in Figure 5. Most often ($n = 15$), a website showed alternative authentication possibilities and directed the user to a direct contact form or email address where they could ask for help if fallback authentication did not work as well. Second most often ($n = 14$) was the exact opposite, where a service did not show any support at all during login; thus, the user's only possibility is to cancel the login attempt and look for help somewhere else. Table 1 provides an overview of the types and extent of support provided by various services during login.

Easiest Option is to use an Alternative Method If the user implemented a backup method (e.g., alternative email or phone number) or has access to backup codes, this is a simple and fast solution to regain access. During login, 50 services suggested using an alternative to the primary second factor. Interestingly, 16 further services generally offered backup methods during the setup but did not mention them during login.

Websites Could Have Directed us to a More Helpful Site Part of the task description was that the researcher had no access to the backup codes, so they looked for solutions outside of the login screen if the login screen was not helpful. Over half (52.6%) of the websites either did not link to any help at all or directed the user to a general help page. For these cases, we additionally tried to find a specific site that explained the procedure a user has to follow when losing access to the second factor. Interestingly, most websites that did not provide specific help when logging in offer a specific FAQ page related to the topic (90.2% of 41). This is especially striking for the 14 websites not supporting the user at all during login: All of them have a specific subpage explaining at least partially what to do.

Tales From the Login Land An existing support site was no guarantee for a goal-oriented process. There were five cases where the suggested or obvious procedure was not helpful at all. In three of those cases, we were stuck in an infinite loop, e.g., because the login screen directed to a help site with a button labeled "Account Recovery;" however, when clicking this, we were directed back to the initial login screen.

We also encountered that the linked support page was only available in the language of the sites' country we could not understand (without a translator). Please note that the rest of the site was available in other languages.

Apart from those five, one page did not provide us with a link to their support until having received a timeout for receiving the code via SMS. In case of a lost phone, this makes the search for help unnecessarily confusing.

4.3.2 Regaining Access

As shown in Figure 6, we were able to regain full access for 41 (52.6%) of the accounts. In nine additional cases, full access most likely would have been possible if we used the account properly and could provide the support with account information, such as banking details, that we did not add to the test account. In one case, the uploaded ID was not accepted, but we received no detailed feedback. We assume that more trials might have given full access.

"Backup Codes are the ONLY Possibility to Access the Account" As mentioned in Section 4.2, ten services explicitly said that users would lose access to their account if they had neither their device nor any backup code. Yet, on six of those, we gained full access after contacting the support. There were essentially two different cases. 1) Three requested details about the account owner or the account like a copy of an identity document, payment details, the address, or the

Total No. Services	No. where better FAQ exists	**Link to support**	Total No. Services	No. where better FAQ exists
15	-	**Direct Form**	5	-
6	-	Specific FAQ	1	-
6	4	General FAQ	1	1
3	2	Unusable	2	0
10	5	But UI support	5	5
10	6	Nothing	14	14

Use of backup suggested

Use of backup not suggested

Link to support given

No link to support given

Table 1: The table depicts the level of support a user gets during the login if they cannot access their second factor. The colors indicate whether a service a) suggests using a backup (e.g., sending the code via mail instead of SMS) and b) if a service provides the user with a link to any support. We also note how many services have a specific information site for 2FA recovery despite not linking to it on the login screen. The most common level of help was given by 15 services: Suggesting to use a backup and linking to a direct form to contact the services' support. On the other hand, 14 services do not support the user at all during login.

current IP address.[2] 2) For three other services, we gained access very easily. One support gave us access after answering a security question. As the researcher was not sure what the answer was, we got a hint after a close-to-correct attempt: ("your answer is close to being correct but is just missing something additional").

Obscure Procedures In the case of a meeting platform, we were asked for our personal meeting-ID. As we did not use the account, we did not store this anywhere and were thus not able to provide it. Interestingly, after disclaiming that we did not have access to this, the second factor was disabled anyway. Since we did not investigate the easiness of accessing the account specifically from an attacker's view, it is up to future work to understand how often information that is asked for is indeed not needed. On another website, we only had to send an email without providing further information, which resulted in us regaining access to the service. We assume, or hope, that this website has internal metrics that allowed them to judge our request. In any case, they did not communicate with us beforehand or even afterward.

4.3.3 Ways to Recover Accounts

We gained full access to our account on 41 services. We could simply receive the 2FA code via email in six of those cases. For the remaining services, we had to contact the services' support.

In the following, we give an overview of what information we had to provide to gain access. We identified five categories of information and evidence services that were asked for proof of ownership during the recovery:

- Personal information, such as name or address.
- Uploading an identity document .
- Basic account information, such as the username or payment details.
- Extended account information, such as information about the last purchase or the date the account was created.[3]
- The need to access the email address used to set up the account.

In general, we saw 17 different combinations of these categories for the 41 accounts we could access. Most commonly ($n = 7$), we were asked for basic account information and needed access to the email address linked to the account. On three services, accessing the account was very easy, as we only needed to provide the service with the email address used for the account, for which we wanted to deactivate 2FA. These services sent a confirmation via mail, but we did not need to react to it with, e.g., clicking a link.

Wait Time Seven services included a wait time for security purposes, meaning they would send a note to the email associated with the account. If they did not receive any negative feedback within a certain time, they would proceed to either delete the account or grant access to it. This waiting time ranged from 1 to 30 days.

No Access but Receiving Additional Help On 37 services, we were not able to regain access to the account. While most mentioned that they could not help us, four provided some

[2] We were in contact with the support via email and believe the IP was used to compare it with IP addresses that were previously used to access the account.

[3] While it was easy to provide the account creation date in our scenario, this question could be tough for users who have had their accounts for many years.

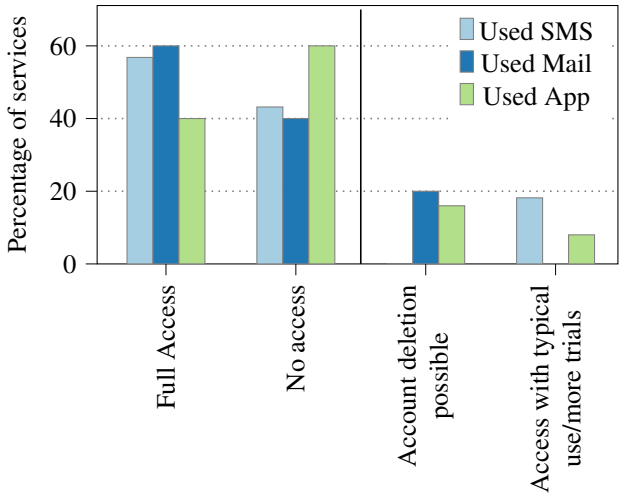


Figure 6: Overview of our results for the recovery process. It shows the percentages of services we either got full or no access to. Two accounts could be recovered using Authy. The right part shows the "no access" category in more detail. For some accounts, an account deletion was possible, on others, we assume we could have logged in with a more realistic setup. The percentages are grouped depending on the second factor we set up before. In the figure, we omitted the two apps where we used passwords as a second factor. In one of those cases, we could have deleted the account; in the other, we could have gained access after a security wait time but without restoring the data that was not backed up.

level of additional help, e.g., they recommended contacting our network provider to receive a new SIM card, which would fix the issue of not receiving SMS codes.

Email as a Second Factor We could recover three of the five services where we used our email as the second factor. To accomplish this, we always had to present the service with another email address. Apart from that, the services differed. In one case, it was sufficient to wait for a month. In the other two cases, we had to provide personal and account information or even upload our ID.

4.3.4 Summary of the Recovery Phase (RQ2+RQ3)

We found that almost the same number of services offered the user no support at all during login as services that gave the maximum possible support by presenting the user with the opportunity of using a fallback authentication as well as a direct contact possibility. Between these two extremes, we saw a lot of different approaches varying in helpfulness. Regarding account recovery, we were successful in regaining access to 41 accounts. However, there were cases where no additional info other than knowing the email address was necessary to disable 2FA.

5 Discussion

We conducted 78 expert reviews to study the setup of 2FA and account recovery on popular services that offer 2FA. In all 78 cases, we were able to successfully set up a second factor. However, our main interest was account recovery. We focused on the information a user was given during setup and the information and guidance these services offered in case the second factor was lost. We could recover access to 41 services without the second factor and backup codes. In general, we found the usability of the setup and recovery process to be lacking in many basic aspects. We discuss themes we saw and make suggestions to practitioners and the research community.

5.1 The User is Often Left Alone

Based on related literature that often mentioned fear of losing the second factor as a reason for not adopting 2FA [10, 25, 35], we phrased our research questions and were especially interested in how services communicate the mitigation and consequences of the loss of the second factor. Both during 2FA setup and recovery, we ran into situations where we only faced vague information or no help at all. During login, 16 services did not inform the user that backup codes can be used instead of codes generated by an authenticator app or sent via SMS, even though they existed. Services that communicated a potential loss of the second factor during setup and that offered backups often avoided statements about accessing the account without the second factor. Only ten services clearly stated that a certain backup would be the only way to gain access. Most other services framed consequences ambiguously, e.g., by stating users "might lose access."

When searching for help at the login screen in the event of a lost second factor, many websites linked to no specific help page, even though one would have existed. All of these problems go against the tenth principle of Nielsen's usability heuristics (help and documentation) [32], and we strongly advise website architects to resolve these easily-fixable issues by adding links to already existing documentation or communicating the possibility of using backup codes during login.

The issue of lacking information is also documented by related work concerning account remediation [28, 31].

5.2 There is no Common Workflow...

We could not identify a common workflow to add a second factor or to recover an account across the different services. This affected all parts of the process: the communication of possibilities for backups, the way a website communicates the consequences of loss, using unified terms, or what information a user needs to provide to recover the account. Currently, a user cannot infer from their experiences from one website to

another. With this, the fourth usability heuristic by Nielsen is violated (consistency and standards) [32].

In our view, this is a problematic situation, as 2FA in itself is a general technical measure to increase account security, and its' usage is likely to increase in the near future.

The origin of this heterogeneity is unclear. Maybe there has not yet been enough time elapsed for a best practice to evolve that everyone copies and can easily adopt. If this is the case, this is also an excellent opportunity to develop a best practice example and provide a fast, secure, and empirical evidence-based solution.

5.2.1 ... not Even Within Services

In addition to the above, many services are not even consistent within themselves. We found one example where 2FA was set up using SMS codes, but the code was sent via email during our login attempt. In another case, a button for "account recovery" existed in the FAQ but linked to the login screen. All of the websites that did not help the user at all during login had a subpage in their support section that explained what to do when the second factor is lost. Similarly, several of those websites that linked to a general FAQ could have linked to a more specific one, making the process much more user-friendly. On some websites, consequences of loss are communicated clearly within such help pages, and several also point to actions that can be done to prevent account lockout. Yet, this is barely mentioned during setup. We believe it is unreasonable to assume that users first look for this specific information on help pages before or after deciding to activate 2FA. Even if it is offered during set-up, users might click through the information, but it is more likely to be seen than if users have to actively look for it (and know-how, too). Fortunately, this is often an easy fix, and we are currently in the process of contacting the affected services to inform them.

5.3 Insufficient Support Structures

The services we used for our research are all popular services. Thus, they handle a lot of traffic and many users. Support on these services is often handled by a bot (chat or phone), and direct human-to-human support was often harder to find. This is fairly common and is likely driven by cost-cutting reasons.

However, depending on the service, it can be very detrimental and stressful to be locked out. We believe that the support structure of many of the services we analyzed does not fulfill the users' needs. We saw cases where support was only available for logged-in users or users who selected a paid product (with no real help available for those using a free version). One website did not offer any help article, but we found a community forum in which frustrated users explained what answers had to be given to the phone bot to end up with a human who could disable 2FA.

Depending on the kind of service, it might be reasonable from the website's perspective not to invest much into recovery procedures, especially in the case of unpaid accounts. Yet, we believe that any account can have a huge value, depending on who is using it for what, and that most users who turn on 2FA voluntarily do see value in their account.

From a usability perspective, we think there should be a dedicated channel for account-related cases. Or, if no dedicated channel is possible, services should at least provide upfront and transparent information on what can be done in such situations.

5.4 Summary: Recommendations for Websites

Summarizing Sections 5.1 to 5.3, we give the following recommendations to website providers:

1. Internal consistency and clear communication during login on what is possible and what is not. E.g., if backup codes exist, the website should mention them as an alternative. If an account cannot be recovered at all, this information should be clearly stated.

2. Services should provide some help during login, similar to the 'forgot password'-link.

3. This help should be as specific as possible. E.g., if the website offers a specific help page explaining how the account can be recovered, this should be directly linked. Preferably, every website had a specific form for this problem, so users could directly contact support.

5.5 Various (and Obscure) Options for Access

In our sample, it was rare to find cases where it was explicitly stated what information a user needs to regain access to their account in the absence of a backup. During recovery, we noticed situations in which access was accomplished very easily, and it was unclear if any technical measures were implemented that checked for the legitimacy of a request to disable 2FA (e.g., using the IP address). Results from Gavazzi et al. [15] indicate that only 22% of their investigated websites block suspicious login attempts, so if this also applies to the aforementioned sites, an attacker might easily get access to the account even if they only know the password.

This is a problem from both usability and security perspectives: The user has no possibility to assess whether the account is really as secure as hoped, i.e., how easy it is for an attacker to disable 2FA. We think when it is not communicated beforehand how access can be granted, users could get a false sense of security.

Similarly, in six cases, we were able to receive the code via email instead of SMS or the authenticator app. If, in these cases, the password can also be reset via email, an attacker

would not need any extra effort to get access as soon as they have control over the email address.

Future work should investigate whether and how users benefit from clear information about 2FA deactivation during or after setting up a second factor.

One solution for a service to make sure a request to disable 2FA is legitimate, also used by 1Password [14], is to combine several proofs of ownership, e.g., requesting access to the email address and also asking for extended account information (knowledge-based challenges). Doerfler et al. [11] studied several of such challenges individually, finding that only 13% of users in their data set were able to recall their account creation date and only 22% could answer their security question.

It remains to be investigated how usable and secure combinations of different challenges are and whether an optimal recovery procedure can be found.

5.6 Who Should be Responsible for Recovery?

We found many opportunities to make 2FA on services much more usable but found this directly connected to the question of who is or should be responsible for a successful recovery.

Most services provide the possibility to recover from lost passwords, so we believe many users might transfer this practice to 2FA.

Yet, we found that while some work has been conducted on how well different fallback authentication mechanisms work (e.g., [11, 29]), we currently do not know what the user's expectations are. Similarly, there is a lack of literature about how website owners and operators see this. It seems that the implicit mindset is that users are responsible for protecting access, including the backup. In any case, we think the easiest mitigation is currently on the side of the services. Transparency could resolve a lot of potential confusion without adding any obvious disadvantages. Golla et al. [20] found that telling people they are responsible for their accounts' security leads to higher adoption of 2FA. The same might apply to backups if the services clearly communicated the consequences.

Authenticator Apps Some authenticator apps provide backup possibilities, yet most rely on passwords, SMS, or emails [18]. Any backup possibilities offered by authenticator apps are currently not part of services' communication, and the Google authenticator is the app most commonly mentioned or recommended by the services ($n = 27$). Interestingly, at the time of the study, Google authenticator only provided one backup possibility, namely a manual QR code export [17, 18]. Since April 2023, Google Authenticator can be synchronized with the users' Google account [4].

Third Parties / Delegated Account Recovery Handling identities connected to user accounts can be challenging. We encountered one website that outsourced this. The website offered to start a recovery over PayPal if a PayPal account was connected to the account. Basically, this follows the idea of SSO. Only a handful of services are responsible for handling the identity. What worked for this website may not work for others, but it opens the question of whether one (or a few) single instances that provide 2FA should also handle the backup and recovery process. In our sample, some services referred to Authy for the recovery process. While, from a usability perspective, this worked well for us, Gilsenan et al. [18] note that Authy solely relies on SMS OTP during recovery. The authors also found several security and privacy issues [16].

5.7 Limitations

Our work has to be interpreted in light of the following limitations:

We focused our analysis on high-traffic websites, so we cannot generalize our results to less popular ones. Yet, we were able to identify issues on these top websites already and believe that administrators and web designers of less popular services can benefit from our results as well.

Not all services support identical second factors (see Section 4.1), but the recovery protocol of services might be influenced depending on the used second factor. We deal with this limitation by giving extra care when comparing the services and pointing to this difference in the results.

Access to some of the services is typically done through the smartphone app. Whenever possible, we used a browser. Thus, it might be possible that the app's interface, including links to the support, differs from the browser version.

Every recovery was made using the same IP that was also used for setup. However, we reinstalled the browser. We cannot estimate how many services checked such metadata before granting access to the account. Additionally, by reinstalling the browser, we chose a tougher scenario than many users would most likely face. We opted for this to capture the lower bound. Similarly, we noticed services that advised us to use a still-logged-in device to disable 2FA. It is up to future work to analyze this in more detail.

We used the accounts only for a short time and only for testing the recovery itself, which comes with further limitations:

- Some services rely on data that is stored within the account to be able to grant access after losing the second factor, e.g., by asking for personal data such as the address or banking details or for order numbers from previous transactions. As we did not add any information, we could not always mimic the whole recovery process. With the empty accounts, we also see the possibility that people working in the support might not have protected the account as much as they would have

with a regularly used account. This is especially critical for services where we were in contact with humans (see Section 4.3.2).

- Some websites periodically ask their users to review and confirm their recovery settings, but we could and did not investigate this feature.

- Some security features might be bound to the users' location or the time they have already used the account. Such details are not captured in our study.

5.8 Future Work

We encountered services that only asked for very basic information to grant us access to the accounts. Similar to as it has been done with security questions [40], it should be studied from an attacker's point of view how easy it would be to get access in such cases.

Two-factor authentication is not the only case where well-designed recovery processes are important. The rise of passwordless authentication is a quite recent example where these processes become crucial, a challenge that future work needs to address.

Due to the many serious issues that we discovered during our 78 expert reviews, we believe that currently, a study that evaluates the usability of account recovery for a lost second factor in a user study would not add much more insight. We are currently in the process of informing the services for which we identified issues.

6 Conclusion

In this study, we aimed to understand how popular services guide their users through the setup of 2FA and the recovery process when the second factor is lost.

We conducted expert reviews on 78 services, analyzing their approach to inform users of possible risks of 2FA, the availability of backup options, and how well users are supported if they cannot access the second factor during login.

Our results revealed that services do not seem to follow a standardized practice for 2FA setup or recovery, and the level of support provided varies greatly among them.

Our findings indicate that only a small percentage of services communicate the importance of a fallback. Additionally, some services do not provide any help during the recovery process, and users are left on their own to solve the issue. These findings suggest that there is room for improvement. Many services could benefit from establishing a more standardized approach to 2FA setup and recovery to ensure convenience for their users without sacrificing security.

Acknowledgments

We thank the Werner Siemens-Stiftung (WSS) for their generous support of this project and our anonymous reviewers for their help and feedback.

References

[1] Jacob Abbott and Sameer Patil. How Mandatory Second Factor Affects the Authentication User Experience. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13. Association for Computing Machinery, New York, NY, USA, April 2020.

[2] Claudia Ziegler Acemyan, Philip Kortum, Jeffrey Xiong, and Dan S. Wallach. 2FA Might Be Secure, But It's Not Usable: A Summative Usability Assessment of Google's Two-factor Authentication (2FA) Methods. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 62, pages 1141–1145, September 2018.

[3] Bitkom. Gestohlen oder verloren: Vier von zehn Personen ist schon mal das Handy abhandengekommen. https://www.bitkom.org/Presse/Presseinformation/Gestohlen-oder-verloren-Vier-von-zehn-Personen-ist-schon-mal-das-Handy-abhandengekommen. Accessed: June 08, 2022.

[4] Christiaan Brand. Google Online Security Blog: Google Authenticator now supports Google Account synchronization. https://security.googleblog.com/2023/04/google-authenticator-now-supports.html, 2023. Accessed: June 08, 2023.

[5] Karoline Busse, Julia Schäfer, and Matthew Smith. Replication: No One Can Hack My Mind Revisiting a Study on Expert and Non-Expert Security Practices and Advice. In *Proceedings of Symposium on Usable Privacy and Security*. USENIX Association, 2019.

[6] Andy C. How To Avoid Losing Your Phone. https://www.mobiles.co.uk/blog/how-to-avoid-losing-your-phone/. Accessed: June 08, 2023.

[7] Dave Childers. State of the auth 2021. https://duo.com/assets/ebooks/state-of-the-auth-2021.pdf, 2021. Accessed: June 08, 2023.

[8] Jessica Colnago, Summer Devlin, Maggie Oates, Chelse Swoopes, Lujo Bauer, Lorrie Cranor, and Nicolas Christin. "It's not actually that horrible": Exploring Adoption of Two-Factor Authentication at a University. In *Proceedings of the 2018 CHI Conference on Human*

Factors in Computing Systems, pages 1–11, Montreal QC Canada, April 2018. ACM.

[9] Sanchari Das, Andrew Dingman, and L. Jean Camp. Why Johnny Doesn't Use Two Factor A Two-Phase Usability Study of the FIDO U2F Security Key. In *Financial Cryptography and Data Security*, volume 10957, pages 160–179. Springer Berlin Heidelberg, Berlin, Heidelberg, 2018.

[10] Sanchari Das, Gianpaolo Russo, Andrew C. Dingman, Jayati Dev, Olivia Kenny, and L. Jean Camp. A qualitative study on usability and acceptability of Yubico security key. In *Proceedings of the 7th Workshop on Socio-Technical Aspects in Security and Trust*, STAST '17, pages 28–39, New York, NY, USA, December 2018. Association for Computing Machinery.

[11] Periwinkle Doerfler, Kurt Thomas, Maija Marincenko, Juri Ranieri, Yu Jiang, Angelika Moscicki, and Damon McCoy. Evaluating Login Challenges as a Defense Against Account Takeover. In *The World Wide Web Conference on - WWW '19*, pages 372–382, San Francisco, CA, USA, 2019. ACM Press.

[12] Jonathan Dutson, Danny Allen, Dennis Eggett, and Kent Seamons. Don't Punish all of us: Measuring User Attitudes about Two-Factor Authentication. In *2019 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, pages 119–128, June 2019.

[13] Florian M Farke, Lennart Lorenz, Theodor Schnitzler, Philipp Markert, and Markus Dürmuth. "You still use the password after all" – Exploring FIDO2 Security Keys in a Small Company. In *Sixteenth Symposium on Usable Privacy and Security*, page 18, 2020.

[14] Pilar Garcia. 10 - Pilar Garcia - Who are you again? Verifying user access rights in an encryption based system. https://www.youtube.com/watch?v=JeV_rop5nmQ, 2019. Accessed: June 08, 2023.

[15] Anthony Gavazzi, Ryan Williams, and Engin Kirda. A Study of Multi-Factor and Risk-Based Authentication Availability. In *32st USENIX Security Symposium (USENIX Security 23)*, 2023.

[16] Conor Gilsenan and Noura Alomar. On Conducting Systematic Security and Privacy Analyses of TOTP 2FA Apps. In *Who Are You?! Adventures in Authentication Workshop*, WAY '20, pages 1–6, Virtual Conference, August 2020.

[17] Conor Gilsenan, Noura Alomar, Andrew Huang, and Serge Egelman. Decentralized backup and recovery of TOTP secrets. In *Proceedings of the 7th Symposium on Hot Topics in the Science of Security*, pages 1–2, Lawrence Kansas, September 2020. ACM.

[18] Conor Gilsenan, Fuzail Shakir, Noura Alomar, and Serge Egelman. Security and Privacy Failures in Popular 2FA Apps. In *32st USENIX Security Symposium (USENIX Security 23)*, 2023.

[19] GitHub. Top-500 npm package maintainers now require 2FA. https://github.blog/changelog/2022-05-31-top-500-npm-package-maintainers-now-require-2fa/. Accessed: June 08, 2023.

[20] Maximilian Golla, Grant Ho, Marika Lohmus, Monica Pulluri, and Elissa M Redmiles. Driving 2FA Adoption at Scale: Optimizing Two-Factor Authentication Notification Design Patterns. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 109–126. USENIX Association, August 2021.

[21] Paul A Grassi, James L Fenton, Elaine M Newton, Ray A Perlner, Andrew R Regenscheid, William E Burr, Justin P Richer, Naomi B Lefkovitz, Jamie M Danker, Yee-Yin Choong, Kristen K Greene, and Mary F Theofanos. Digital identity guidelines: authentication and lifecycle management. Technical Report NIST SP 800-63b, National Institute of Standards and Technology, Gaithersburg, MD, June 2017.

[22] Beatriz Henríquez. Mobile Theft and Loss Report - 2020/2021 Edition | Prey Blog. https://preyproject.com/blog/mobile-theft-and-loss-report-2020-2021-edition. Accessed: June 08, 2023.

[23] Andy Homan. 44% of people lose their mobile. https://nuttag.com.au/blogs/news/44-of-people-loose-their-mobile. Accessed: June 08, 2023.

[24] Iulia Ion, Rob Reeder, and Sunny Consolvo. "... no one can hack my mind": Comparing Expert and Non-Expert Security Practices. In *Eleventh Symposium On Usable Privacy and Security (SOUPS 2015)*, pages 327–346, 2015.

[25] Kat Krol, Eleni Philippou, Emiliano De Cristofaro, and M. Angela Sasse. "They brought in the horrible key ring thing!" Analysing the Usability of Two-Factor Authentication in UK Online Banking. In *Proceedings 2015 Workshop on Usable Security*, San Diego, CA, 2015. Internet Society.

[26] Yue Li, Haining Wang, and Kun Sun. Email as a Master Key: Analyzing Account Recovery in the Wild. In *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications*, pages 1646–1654, Honolulu, HI, April 2018. IEEE.

[27] Lookout. PHONE THEFT IN AMERICA. https://transition.fcc.gov/cgb/events/Lookout-phone-theft-in-america.pdf. Accessed: June 08, 2023.

[28] Philipp Markert, Andrick Adhikari, and Sanchari Das. A Transcontinental Analysis of Account Remediation Protocols of Popular Websites. In *Proceedings 2023 Symposium on Usable Security*. Internet Society, 2023.

[29] Philipp Markert, Maximilian Golla, Elizabeth Stobert, and Markus Dürmuth. Work in Progress: A Comparative Long-Term Study of Fallback Authentication. In *Proceedings 2019 Workshop on Usable Security*, San Diego, CA, 2019. Internet Society.

[30] Karola Marky, Kirill Ragozin, George Chernyshov, Andrii Matviienko, Martin Schmitz, Max Mühlhäuser, Chloe Eghtebas, and Kai Kunze. "Nah, it's just annoying!" A Deep Dive into User Perceptions of Two-Factor Authentication. *ACM Transactions on Computer-Human Interaction*, December 2021.

[31] Lorenzo Neil, Elijah Bouma-Sims, Evan Lafontaine, Yasemin Acar, and Bradley Reaves. Investigating Web Service Account Remediation Advice. In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*, pages 359–376. USENIX Association, August 2021.

[32] Jakob Nielsen. 10 Usability Heuristics for User Interface Design. https://www.nngroup.com/articles/ten-usability-heuristics/, 2021. Accessed: June 08, 2023.

[33] NIST. NIST Special Publication 800-63: Digital Identity Guidelines - FAQ. https://pages.nist.gov/800-63-FAQ/#q-b11, 2022. Accessed: June 08, 2023.

[34] European Parliament and Council of the European Union. DIRECTIVE (EU) 2015/2366 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32015L2366. Accessed: June 08, 2023.

[35] Jeunese Payne, Graeme Jenkinson, Frank Stajano, M. Angela Sasse, and Max Spencer. Responsibility and Tangible Security: Towards a Theory of User Acceptance of Security Tokens. In *Proceedings 2016 Workshop on Usable Security*, San Diego, CA, 2016. Internet Society.

[36] Victor Le Pochat, Tom Van Goethem, Samaneh Tajalizadehkhoob, Maciej Korczyński, and Wouter Joosen. Tranco: A Research-Oriented Top Sites Ranking Hardened Against Manipulation. In *Proceedings 2019 Network and Distributed System Security Symposium*, 2019.

[37] Nils Quermann, Marian Harbach, and Markus Dürmuth. The State of User Authentication in the Wild. In *Who Are You?! Adventures in Authentication Workshop (WAY) 2018*, Baltimore, MD, USA, August 2018.

[38] Ken Reese, Trevor Smith, Jonathan Dutson, Jonathan Armknecht, Jacob Cameron, and Kent Seamons. A Usability Study of Five Two-Factor Authentication Methods. In *Fifteenth Symposium on Usable Privacy and Security (SOUPS 2019)*, pages 357–370, Santa Clara, CA, August 2019. USENIX Association.

[39] Joshua Reynolds, Trevor Smith, Ken Reese, Luke Dickinson, Scott Ruoti, and Kent Seamons. A Tale of Two Studies: The Best and Worst of YubiKey Usability. In *2018 IEEE Symposium on Security and Privacy (SP)*, pages 872–888, San Francisco, CA, May 2018. IEEE.

[40] Stuart Schechter, A.J. Brush, and Serge Egelman. It's no secret: Measuring the security and reliability of authentication via 'secret' questions. In *Proceedings of the 2009 IEEE symposium on security and privacy*. IEEE Computer Society, May 2009.

[41] Statista. Most convenient Multi-Factor Authentication (MFA) methods worldwide in 2021. https://www.statista.com/statistics/1303617/convenient-global-mfa-methods/, 2021. Accessed: June 08, 2023.

[42] Tranco. Information on the Tranco list with ID X5KYN. https://tranco-list.eu/list/X5KYN/1000000, 2022. Accessed: June 08, 2023.

[43] Jake Weidman and Jens Grossklags. I Like It, but I Hate It: Employee Perceptions Towards an Institutional Transition to BYOD Second-Factor Authentication. In *Proceedings of the 33rd Annual Computer Security Applications Conference*, pages 212–224, Orlando FL USA, December 2017. ACM.

[44] Google Authenticator. https://play.google.com/store/apps/details?id=com.google.android.apps.authenticator2. Accessed: June 08, 2023.

A Website Analysis

A.1 Exclusion criteria for services

Services were excluded for the following reasons:

- **Security or Accessibility:** websites flagged as dangerous by Google Safe Browsing, URLs not belonging to a DNS server or are unreachable
- **Content**: adult entertainment websites or illicit content

- **Shared login:** sites belong to the same domain as a previously listed site and having shared accounts (e.g., Google and Youtube)
- **Language**: sites that don't provide an English or German interface
- **Payment**: requiring payment details for account setup. If a free short-term trial was available, we used this opportunity.
- **Specific user group:** requiring owning a product for account setup, accounts requiring the user to be in a specific region outside of Germany, sites restricted to specific users (e.g., university websites, accessible only by students and faculty members)
- **Additional steps:** requiring in-person interactions

Service	Second Factor	Setup				Recovery	
		Hint	Backup	Force	Shown in...	Link to Support	Suggests Using Backup
AOL.com	SMS	●	●	●	▭	○ But UI Support	●
Abusix.com	App	●	●	○	⚙	○ But UI Support	●
Adobe.com	SMS	●	●	●	▭	○ But UI Support	●
Amazon.com	SMS	○	●	○	⚙	● Direct Form	○
Apple.com	SMS	○	○	○	-	● Direct Form	○
Avast.com	App	●	●	○	⚙	● Unusable	●
Bit.ly	SMS	○	○	○	-	○	○
Booking.com	SMS	○	○	○	-	○ But UI Support	●
CloudDNS.net	App	●	●	○	⚙	○	●
Cloudflare.com	App	●	●	○	⚙	● Direct Form	●
Cloudone.trendmicro.com	App	●	●	●	▭	○ But UI Support	●
DNSmadeeasy.com	App	●	●	○	⚙	● Direct Form	●
Digicert.com	App	○	○	○	-	○ But UI Support	○
Discord.com	App	●	●	○	⚙	○	●
Dropbox.com	SMS	●	●	○	⚙	● Specific FAQ	●
Ebay.com	SMS	●	●	●	▭	○	●
Epicgames.com	SMS	○	●	○	⚙	○	●
Etsy.com	SMS	●	●	○	⚙	○	○
Facebook.com	SMS	●	●	○	⚙	● Direct Form	●
Fastly.net	App	●	●	○	⚙	● General FAQ	●
Fedex.com	SMS	○	○	○	-	● General FAQ	●
Fiverr.com	SMS	○	●	●	▭	● Direct Form	●
Gcore.com	App	○	●	○	▭	○ But UI Support	○
Gandi.net	App	●	●	●	▭	● General FAQ	●
Github.com	SMS	●	●	●	▭	● Direct Form	●
Godaddy.com	SMS	○	●	○	⚙	● Specific FAQ	○
Google.com	SMS	●	●	○	⚙	● Unusable	●
Grammarly.com	SMS	●	●	●	▭	● Direct Form	●
HP.com	App	○	○	○	-	○	●
Herokuapp.com	App	○	●	○	▭	○	○
IlovePDF.com	App	○	○	○	-	● Unusable	○
Indeed.com	SMS	○	○	○	-	○ But UI Support	○
Instagram.com	SMS	○	○	○	-	○ But UI Support	●
Intuit.com	SMS	○	●	○	▭	● Unusable	●
Kaspersky.com	SMS	○	●	○	▭	○ But UI Support	○
Kickstarter.com	SMS	○	○	○	-	○	○
Linkedin.com	SMS	○	○	○	-	● Direct Form	○
Linktr.ee	SMS	○	○	○	-	○	○
Mailchimp.com	SMS	○	○	○	-	○	○
Microsoft.com	Email	●	●	●	▭	● Direct Form	●
MyShopify.com	SMS	●	●	○	⚙	● General FAQ	●
Name.com	App	●	●	○	⚙	● General FAQ	○
No-IP.com	App	●	●	○	⚙	○ But UI Support	●
OK.ru	SMS	○	●	○	▭	● Unusable	○
Onlyfans.com	SMS	○	●	○	▭	○ But UI Support	○
Opera.com	App	●	●	●	▭	○	●
Patreon.com	SMS	●	●	○	⚙	○	●
Paypal.com	SMS	○	●	○	▭	● Direct Form	○
Pinterest.com	SMS	●	●	○	⚙	○	○
Reddit.com	App	●	●	○	⚙	○	●
Ring.com	SMS	○	○	○	-	○	○

Service	Second Factor	Setup				Recovery	
		Hint	Backup	Force	Shown in...	Link to Support	Suggests Using Backup
Roblox.com	Email	●	●	○	⚙	● Specific FAQ	●
Samsung.com	SMS	●	●	○	⚙	● Direct Form	●
Slack.com	SMS	●	●	○	⚙	● Specific FAQ	●
Snapchat.com	SMS	●	●	○	⚙	○	○
Sourceforge.net	App	●	●	○	⚙	○ But UI Support	●
Squarespace.com	App	●	●	○	⚙	● Specific FAQ	●
Steampowered.com	Email	○	○	○	-	● Direct Form	○
Stripe.com	SMS	●	●	○	⚙	● Direct Form	●
Teamviewer.com	App	●	●	●	▭	● General FAQ	●
Telegram.org	Password	●	●	○	⚙	● Direct Form	●
ThemeForest.net	App	●	●	○	⚙	○	○
Tiktok.com	SMS	●	●	●	▭	○	●
Tinyurl.com	App	●	●	○	⚙	○ But UI Support	●
Tradingview.com	SMS	●	●	○	⚙	● Direct Form	●
Trello.com	App	●	●	●	▭	● Direct Form	●
Tumblr.com	SMS	●	●	○	⚙	○	○
Twitch.tv	SMS	●	●	○	⚙	● Specific FAQ	●
Twitter.com	SMS	●	●	○	⚙	● Direct Form	●
Unity3d.com	SMS	●	●	○	⚙	○	●
VK.com	SMS	●	●	○	⚙	○	○
Vimeo.com	Email	○	○	○	-	○	○
Wetransfer.com	App	●	●	●	▭	● Direct Form	●
Whatsapp.com	Password	●	●	○	⚙	● Direct Form	●
Wixsite.com	Email	●	●	○	⚙	○	○
Yahoo.com	SMS	●	●	●	▭	○ But UI Support	●
Zendesk.com	SMS	○	●	○	⚙	● Specific FAQ	●
Zoom.us	SMS	●	●	●	▭	● General FAQ	●

Table 2: Overview of help a user gets during setup and recovery of a second factor. ○: The service does not include the characteristic. ●: The service fulfills the characteristic.
Except for one, all services that offered the user to use a backup during login but did not provide a backup possibility send the code to the email/phone number used to register. One service did not have clear backups but suggested using backup codes during login.

Tangible 2FA – An In-the-Wild Investigation of User-Defined Tangibles for Two-Factor Authentication

Mark Turner[1], Martin Schmitz[2], Morgan Masichi Bierey[1], Mohamed Khamis[1], Karola Marky[1,3]
[1]*University of Glasgow, United Kingdom,* [2]*Saarland University Saarbrücken, Germany,*
[3]*Ruhr-University Bochum, Germany*

Abstract

Although two-factor authentication (2FA) mechanisms can be usable, they poorly integrate into users' daily routines, especially during mobile use. Using tangibles for 2FA is a promising alternative that beneficially combines customisable authentication routines and object geometries, personalisable to each user. Yet, it remains unclear how they integrate into daily routines. In this paper, we first let 226 participants design 2FA tangibles to understand user preferences. Second, we prototyped the most common shapes and performed a one-week long in-the-wild study (N=15) to investigate how 2FA tangibles perform in different environments. We show that most users prefer objects that a) fit in wallets, b) connect to daily items or c) are standalone. Users enjoyed interacting with 2FA tangibles and considered them a viable and more secure alternative. Yet, they voiced concerns on portability. We conclude by an outlook for a real world implementation and distribution of 2FA tangibles addressing user concerns.

1 Introduction

Two-factor authentication (2FA) is has become part of our daily lives, with many services, from banks to major internet players offering the security benefits of 2FA [4, 7]. While these security benefits are undisputed and the early usability problems of the authentication procedure have been mainly resolved by constant improvements (cf. [9, 10]), newer research has shown that a large share of users are still reluctant to use 2FA beyond being forced to do so by their providers [1, 16, 17].

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.
USENIX Symposium on Usable Privacy and Security (SOUPS) 2023.
August 6–8, 2023, Anaheim, CA, USA

The reasons for that lie beyond usability in the users' daily lives, routines, and habits that are interrupted by current 2FA procedures, creating too much so-called *friction* [17, 20], for instance, by taking too long [6, 11, 32] or being not readily available [6, 16, 17]. While previous work has identified these general issues, finding appropriate alternatives that better integrate into users' daily routines and contexts remains an open research challenge.

Among possible alternatives are *tangible* interactions that better integrate into users' individual environments and routines by utilising digital fabrication [21]. Tangibles are physical objects used to manipulate digital information [25]. In the context of 2FA, tangibles can serve as personal user tokens. They either are the authentication factor ownership or form a complete 2FA mechanism. In 2020, 3D-Auth [18] was proposed as a tangible 2FA mechanism. It is based on using 3D-printed tangibles for 2FA that can be customised in terms of colour, and shape interaction and be integrated into other daily items, such as accessories. The 2FA tangible itself embeds a unique conductive structure that can be sensed by touchscreens and encodes the authentication factor ownership. By interacting with the 2FA tangible, users enter a kind of haptic password, e.g., by rotating parts of the tangible. This interaction represents the knowledge authentication factor. Consequently, 3D-Auth offers 2FA in one interface.

The knowledge-based interaction has been demonstrated to have a high memorability since it also leverages muscle memory [18]. What remains unclear though, is what kinds of 2FA tangibles users might want to use for authentication and how these tangibles perform when used on a daily basis. This paper contributes to the space of tangibles for 2FA by investigating the following research questions:

RQ1: *What kinds of 2FA tangibles do users wish to use?* We investigate what kinds of tangibles users wish to use in their daily routines and how they wish to interact with them. For this, we conducted an online study where we let 226 users configure their ideal tangibles.

RQ2: *How do tangibles for 2FA perform in the user's daily lives? What are the obstacles and challenges introduced*

by them? We investigate how 2FA tangibles as a 2FA mechanism that combines ownership and knowledge perform when used daily. For this, we first chose the top three 2FA tangible designs and developed them as prototypes. Second, we conducted an in-the-wild study (N=15) during which participants used the prototypes in their daily life over a whole week. Using short questionnaires during the interaction phase and in-depth interviews, we report and discuss positive aspects and emerging challenges for 2FA tangibles.

Research Contribution. In summary, the main contributions of this paper are:

User-Defined Tangibles: We investigate what kinds of tangible 2FA items users choose in an online study with 226 participants. We show what kinds of interactions users would like to perform for authentication and which shapes, sizes, colours, and further properties they prefer. Participants preferred rather simple geometric shapes, such as cubes or squares with sizes between one and ten centimetres to fit the smartphone screen.
In-the-Wild Investigation: Based on the results of the online study, we designed three tangible prototypes that realise a 2FA mechanism that combines the ownership and knowledge authentication factors. We used these tangibles in an in-the-wild study where 15 participants used the designed tangibles in their lives for a week. Our participants perceived the tangibles as adding a layer of security to their important accounts. In addition, the interactions were mostly perceived as easy-to-use and fun.
User-Centred Design Pipeline: We conclude by proposing a user-centred design pipeline that assists the users in designing 2FA tangibles specifically for their preferred usage environment, security needs, and user preferences. The design pipeline is not limited to standalone 3D-printed tangibles but also considers alternatives with integrated sensors to enable authentication for all kinds of devices, e.g., by using USB connections or NFCs.

2 Background and Related Work

In this section, we present background and related work that our research builds upon.

2FA Realisations. Several realisations of 2FA have been brought to the market or were proposed by related work. The usage of one-time passwords (OTPs), e.g., via text messages, emails, phone calls, or apps nowadays is well-established and used by a plethora of providers. Most of them require a smartphone to somehow access the OTPs. Another smartphone-based option without OTPs are push notifications where users press a button. For those who do not wish to rely on smartphones, OTP generators (e.g., DUO Security Token [29], or Fido U2F [8]) are an alternative. Based on security considerations, the second authentication factor should ideally be on a different device than the main interaction. For instance, if the authentication is done on a laptop, a smartphone or token can be used. If authentication is done on a smartphone, a token or another smartphone would be required. Investigations of real-world usage showed that users frequently use one device for authentication and main interaction [17] which defeats some security benefits of 2FA. Related work also showed that mobile users might not be willing to carry a dedicated OTP generator [17, 31]. Reasons for that were limitations in personalisation [17].

3D-Auth Concept [18]. In this paper, we use the concept 3D-Auth [18] as a basis for our investigation. The concept combines the security benefits of a separate token while also mitigating personalisation issues by allowing users to either choose a custom 3D-printed shape or integrate the token into an everyday object, such as an accessory.
Fabrication: 3D-Auth items are fabricated as follows: they have an internal authentication structure by embedding a capacitive material within insulating plastic. The capacitive material can be detected using a capacitive touchscreen. For this, each 3D-Auth item has a grid of conductive dots at the item's bottom. The conductive dots themselves are not sufficient for detection by the touchscreen, because users have to actively touch the item.
Interaction: Marky et al. [18] present five possible interaction categories for the knowledge-based part. First, users *touch* the item surface on specific spots or perform gestures. Second, users *arrange* one or multiple items on the touchscreen. Third, users *configure* the item by pressing or rotating parts of it. Fourth, users *assemble* a set of multiple items into one item. Finally, users change the item's internal configuration by *augmenting* it with something else (e.g., water or air).
2FA by 3D-Auth: 3D-Auth realises 2FA by splitting the conductive structure into two components: (1) a *static* component that encodes the authentication factor *ownership*. This makes up one subset of the conductive dots that are always sensible by the touchscreen and can be compared to a conductive token; and (2) a *dynamic* component that encodes the authentication factor *knowledge*. Through interaction, the dynamic part of the capacitive authentication structure is transformed in such a way that the change can be detected. This makes up another subset of the conductive dots that are turned into sensible touchpoints by user interaction. This interaction can be compared to a haptic password. Both structures together form the authentication pattern that is sensed by a touchscreen. It is a subset of the conductive dots in the object's bottom.
Security Aspects: 3D-Auth items protect accounts by proving two authentication factors in one item. In comparison to existing standalone authentication tokens that only encode the ownership factor (e.g., non-bio YubiKeys [35]), 3D-Auth offers a higher level of security because the item alone is not enough to impersonate the user due to the dynamic

authentication component. Further, the item cannot easily be replicated (e.g., by observation), because either the 3D printing file is needed, or the item has to be cut into several layers to reveal its entire internal structure. Assuming that an attacker takes over a device (e.g., a smartphone), 3D-Auth offer a higher level of security compared to OTPs via SMS, authenticator apps or other kinds of notifications, because the 3D-Auth item is physically separate from the device. Since this paper focuses on the human perspective of 2FA tangibles, we refer to the original 3D-Auth publication for more in-depth security-related information [18].

Tangible Authentication. Using tangible items for authentication has been proposed before in the scope of single-factor authentication, by using conductive sheets that cover parts of a touchscreen [30, 34] or a Rubik's Cube-like structure where interactions are captured with a camera [22].

Adoption and Usability of 2FA. The reasons for (not) adopting 2FA vary. The main criterion for using 2FA is the security benefit for protecting valuable assets [23, 24]. For instance, the amount of money in accounts impacts 2FA adoption; the more money, the more likely users protect the account with 2FA [23]. Further impacting factors are usability [3, 11, 32], trustworthiness [11], the required cognitive effort [11, 17] and familiarity with 2FA [6, 12, 33]. Abbott and Patil conducted a series of online surveys in a university where 2FA is mandatory [1]. They could not find the mandatory nature to impact the acceptance of 2FA. Instead, motivating users with personalised messages to assist them in adjusting their mental model of 2FA is promising to boost adoption [15].

The usability of different 2FA realisations has been thoroughly researched in the past. Yet, there is no overall consensus since the specific approach and its realisation seem to have a profound impact on usability. Further, two distinct usage phases have to be considered separately: 1) setup phase and 2) authentication phase [2, 26] that we discuss in the following.

The YubiKey [35] is a token that supports several cryptographic protocols, e.g., OpenPGP. It can be connected to a computer via USB or a smartphone via USB-C or NFC and is a possible second factor for 2FA. Participants in several user studies struggled to set up the YubiKey [2, 9, 24, 26]. This was also demonstrated for other tokens [5, 6, 32]. In contrast to tokens, the setup of OTPs by text messages [2, 5, 24], pre-generated OTPs [24], and push notifications [2, 24] was perceived as easier-to-use. Consequently, the setup process of 2FA is crucial; difficult setup procedures might even discourage users from using 2FA at all [2]. However, setup procedures ideally have to be done only once and can be improved.

Considering the authentication phase, several studies demonstrated the usability of OTPs via SMS [2, 11, 16], OTP generators [11, 16], tokens [6, 10, 14] and smartphone apps [2, 11, 16] while pointing out shortcomings that are possible to correct. An example is the research by Das et al., who successfully demonstrated usability improvements of the YubiKey [10]. After an initial study, they refined the YubiKey and demonstrated its improved usability.

Even though the studies mentioned above clearly demonstrated that the authentication phase can be usable, it has also been shown that user experience-related aspects and the user's context play an essential role when adopting 2FA [10, 14, 16, 17, 31]. Participants in studies of tokens, for instance, were not willing to use a token because they feared losing it [10, 14]. Further, participants in studies voiced an unwillingness to set up and carry around single-purpose extra devices [16, 17, 31]. Some OTP generators ran out of battery when needed [17]. While participants could successfully authenticate, the duration of the procedure was perceived as too long [6, 11, 32], especially when doing multiple authentications as part of a daily routine [17].

Summary. 2FA can be usable, but the usability of the setup and authentication phase is not enough. This paper investigates 3D-Auth as an alternative tangible authentication concept. First, we investigated what kind of items users might want to use for authentication. Second, we prototyped the most common shapes and investigated how users interacted with them over a week. Our study considers the participants' contexts and interviews them about 2FA integration in their daily routines.

3 Study I: User-Defined Tangibles

In our first investigation, we wanted to find out what kind of 2FA tangibles users want to use in their daily life, specifically investigating RQ1 (*What kinds of tangibles do users design for 2FA?*). For this, we conducted an online user study with 226 participants. During the study, participants were asked to configure their tangibles as a form of authentication. For this, we implemented a design pipeline, where participants were guided through a design process by picking several tangible properties, as detailed below.

Study Procedure. First, after reading and accepting our consent form, we explained the concept of a 2FA tangible, how it works and possible interactions to the participants in textual form. After a trial run with ten participants, the description texts were refined, and illustrative pictures were added to foster a better understanding. We further added a quiz to the end of the familiarisation part to help participants by testing their understanding. These items served as attention checks in case participants failed them multiple times following the guidelines of Prolific.

Second, participants designed their ideal 2FA tangible following a mock design pipeline. First, they designed the physical appearance of the tangible in free text. Next, they provided specifications on the colour and size of the tangible. Then, they were asked about the desired number of interactions

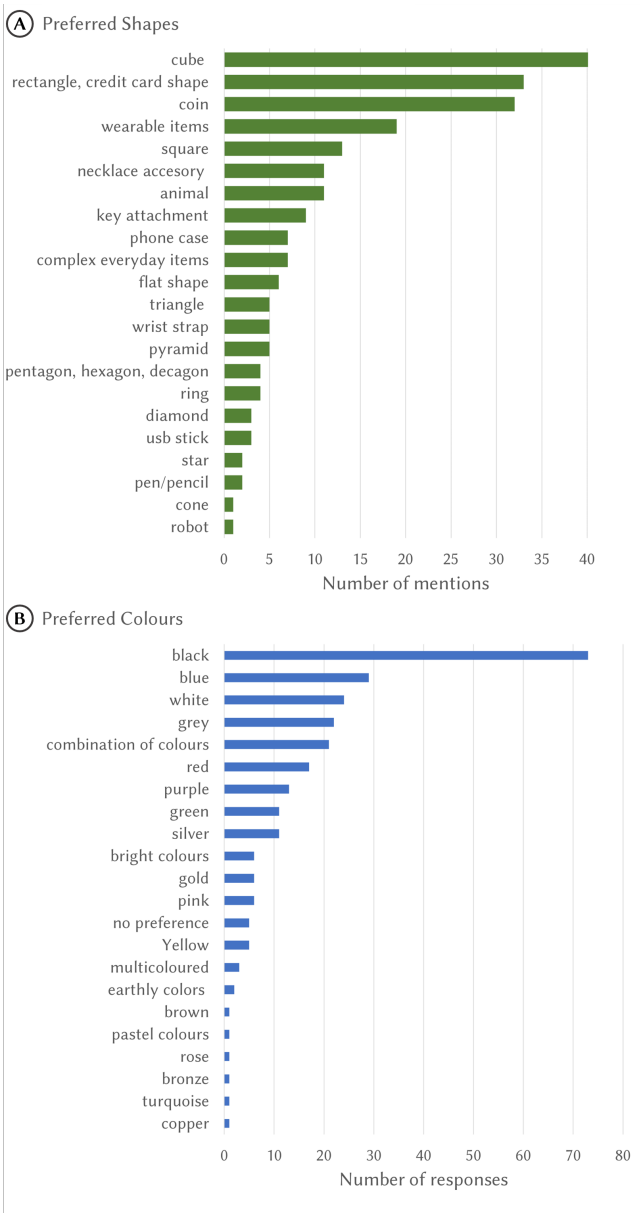


Figure 1: The preferred shapes (a) and colours (b) stated by the participants of the online study.

based on the interaction space with the five interaction categories from Marky et al. [18] and to provide the specific interactions they would like to perform with their tangible. Another attention check was carried out here.

In the last step, participants were asked for demographics including age, gender, origin, and answers to the affinity of technology scale [13]. After the participants completed all question sets, they were redirected to the survey platform for reimbursement.

Recruitment & Participants. We recruited 233 participants using the Prolific online platform[1]. Seven of them failed more than one attention check resulting in 226 valid datasets. Of them, 129 identified as male, 90 as female, and seven identified as self-described. Their mean age was 27 (*min* = 18, *max* = 58, *SD* = 8). Participants were compensated with an hourly rate of an equivalent of 11 US dollars.

Limitations. Like most online studies, our investigation has several limitations among them are wrong self-assessments and biased answers due to social acceptability. It might be challenging for participants to make judgements about configuring 2FA tangibles without the possibility of exploring them physically or having experience interacting with them. The sample might not be representative of the entire population of potential 2FA tangible users. Hence, our results should be validated through future in-depth studies with more heterogeneous samples.

3.1 Results

We analysed the collected designs, colour, and size descriptions by an inductive categorisation approach [19] where two researchers independently grouped the participants' answers into clusters of designs, colours, and sizes. Disagreements were resolved in a review meeting.

Tangible Design. Participants suggested a wide variety of shapes and colours for possible 2FA tangibles (see Figure 1). Possible sizes were clustered into three groups: small (1-3 cm, N=56), medium (4.5-6 cm, N=59), or large objects (8-10 cm, N=42) in quite equal proportions. Instead of giving absolute terms, they linked their favourites to objects they already knew. For example, 25 participants chose size 10 cm as they associated it with a pencil shape they might want to use. Other participants linked their ideas to relative statements, such as *"a thumb"* (P111) or *"a penny"* (P140). Considering the design, three main clusters of tangibles emerged:

(1) Standalone Tangibles: Most participants (N=107) described a standalone tangible for authentication. More specifically, the description from the participants based on the specified shape and properties was a 2FA tangible that is neither connectable nor insertable into another everyday object, such as a wallet. Interestingly, participants preferred rather generic shapes, such as cubes (N=43), squares (N=13), or pyramids (N=5), instead of more complex geometries. The majority of those, who preferred a more complex shape, described animals (N=11). Participants, for instance, envisioned the following 2FA tangibles:*"It could be a cube, that I can play around and rotate the cube so that the side with the internal authentication object would be known only by me."*, **P408**.

[1] https://www.prolific.co, last-accessed 1-February-2023

"I like the idea of having a collection of animal authentication objects. For my taste, probably a cat.", **P118**.

Even though participants were not specifically asked to justify their choices, some of them did so by mentioning portability aspects, such as P149 who wrote *"It would like a pencil but a bit smaller so that I could always carry it with me and use it with no difficulties."*, **P149**.

(2) Wallet-Fit Tangibles: Participants specifically named tangibles that fit into their wallets or pockets (N=65), either credit card-shaped (N=33) or coins (N=32). Participants described these, for instance, as follows:

"[...] the size and shape of a typical credit card.", **P208**.

"It should look like a coin, so it's comfortable to carry in my wallet, pocket, etc.", **P087**.

"A card that displays numbers or codes that could be kept in one's wallet.", **P180**.

(3) Connectable Tangibles: Connectable tangibles are the smallest category (N=54). Those tangibles are somehow connectable to either another daily object in the form of key rings (N=9), and phone cases (N=7) or to the human body as a wearable (N=38). Even though participants were not asked for in-depth justifications, most participants that described connectable tangibles mentioned that they want an object that ideally passes as something that is not linked to authentication:

"'bullet' or 'pendant shaped', such that it could pass as a necklace to someone who didn't know what it was.", **P040**.

"I would prefer a small object, such as a ring or bracelet. Preferably something I can wear.", **P191**.

"The goal for me is to be very discreet, when people see the object they won't know it is an authentication object, so it has to be design/decorative, for example, a phone case, if you touch it at specific points in a specific order it will unlock, but no one will notice.", **P232**.

Interactions. When asked for the preferred number of interactions, 35.71% of participants stated two interactions. A slightly smaller share (34.52%) prefers one interaction. The remainder of the participants stated willingness to use three (16.07%), four (5.95%), five (5.95%), six (1.19%), or ten interactions (0.59%).

In addition, participants were asked to choose interactions for their designed tangible (multiple answers and different interaction combinations were possible). 39.88% of the preferred interactions were touch-based. This was followed by configuration with 24.92% and arrangement (20.23%). Only a few participants preferred assembly (11.43%) and augmentation (2.34%). In 1.17% of the interactions, participants added their descriptions with image recognition, voice interaction, and pinning.

4 Design & Prototype Implementation

In this section, we describe the tangible design process that we followed to develop the 2FA tangibles for Study II. While we initially considered using the 3D-Auth items presented in the literature [18], those did not match the user preferences voiced in Study I. Therefore, we designed a new set of 2FA tangibles that is optimised for mobile usage.

Designing 2FA Tangibles. We used the data set collected in Study I as a basis for designing tangible for Study II. First, we filtered out designs without a flat surface or those that were too small to embed the capacitive material. Then, we developed the first set of candidates based on the interactions preferred in the online study. We filtered out the interactions augmentation and assembly because the majority of participants had concerns about using them on the go. Augmentation might be difficult due to the need for water or some other external media. Assembled tangibles result in more individual objects that might be lost. Finally, we matched the list of candidates with the three tangible categories from the online study to propose several candidates for each category.

All tangibles were designed to fit into pockets, purses and wallets. The static authentication structure was a square shape with an embedded dot structure for each tangible, representing the ownership factor. The dynamic parts and designs are as follows (see also Fig. 2), each interaction serving as the knowledge factor:

1) Wallet Category: Credit card with touch interaction. This tangible has the approximate dimensions of a standard credit card (85mm $\times$ 55mm), with the exception of the thickness, which was made larger (3mm) to ensure that no touches on non-conductive material were registered. To authenticate with the tangible, users place the item on their phone and perform a sliding motion over a circle of ten dots printed with the conductive material. First, the users start touching the top dot and then clockwise to the fifth dot (180°). Then, they move anti-clockwise to the second dot to the left (270°). Finally, the user touches the square structure. This was to mimic using a safe lock.

2) Standalone Category: Cube with arrangement interaction. This tangible has the style of a die (20mm $\times$ 20mm $\times$ 20mm), which many people surveyed suggested in their responses, with each side containing a different number of pips of conductive material. The side of the die that would usually have the number one had the square that encodes the ownership factor. To authenticate, the users touched the smartphone with different sides of the item following the sequence four, one (ownership factor), four, and two. Each of the numbers denotes the number of dots on the respective die side.

3) Connectable Category: Key-chain with configuration interaction. This tangible is a combination lock, with ten possible digits for each layer and three total layers, each assembled onto a central axis. Each layer is 30mm × 30mm × 5mm, with the central axis 23mm tall. To authenticate, users align the three layers matching the number sequence one, three, and seven, similar to the way one would interact with a combination lock, and then touch the item to the phone. Additionally, the central axis was created with conductive material to serve as a 'control', allowing for the phone screen to read unsuccessful attempts.

Prototypical Authentication App. To facilitate the collection of information and allow users to experience performing login authentications with a 2FA tangible, we created a mock authentication app (see Appendix A.4.) for Android that simulated the experience of unlocking a remote account (e.g., emails or online banking) which might currently be unlocked via an OTP received, for instance, by SMS or authenticator app. For this, we implemented the following functionality that also serves as the basis for Study II: First, the app offers a tutorial for each tangible to allow users to learn without requiring a demonstration. Second, the app recognises the interactions using each of the 2FA tangibles to provide a proper authentication experience. If participants could not authenticate, the app allows skipping the process. After the authentication or authentication skip, the app prompts the participants to answer a short survey consisting of different questions depending on the authentication terminal state. Third, the app collects the required data and transmits it to our cloud database via Firebase messaging. Forth, the app receives notifications, so we can nudge participants to authenticate throughout the day. The notification can be snoozed to serve as a reminder to participate in the study at least once a day.

We envision this concept as part of an app that requires login via 2FA where no switch of app or device is needed, hence the mock authentication app was intended to allow a simulation of the need for authentication during the day.

5 Study II: 2FA Tangibles in the Wild

Based on the results of Study I, we conducted a follow-up study to investigate RQ2 (*How do tangibles for 2FA perform in the user's daily lives? What are obstacles and challenges introduced by such novel tangibles?*) For this, the tangibles detailed prior were distributed to 15 participants and used for one week to unlock a remote account.

Collected Data. The data collected during the study by the app fell into two categories: First, *authentication data*, consisting of the time taken to complete the authentication, the timestamp the authentication took place, the number of attempts required, whether the authentication was successful, whether the user skipped the authentication and the user's participant number. Second, the app collected *survey data*. This differed slightly in wording depending on whether the user succeeded, failed, or skipped. Each survey collected information on the user's location when they performed the authentication (multiple choice), what issues the users had, if any (multiple choice), whether they would like to perform this authentication in a similar setting (single choice) and why (open text), and further feedback (open text, optional). All survey questions are provided in Appendix A.1.

Study Procedure. The study consisted of an initial meeting, a one-week usage period (between one and three authentications per day), and an exit meeting including an interview. During the initial meeting, participants were met either in-person or online, with 15 minutes allocated:

First, participants were informed about the concept of 2FA tangibles, and what they would be required to do over the week. Then, informed consent was obtained.

Second, each participant was assigned one tangible at random[2], and installed the study app on their device. The information for the first start of the app (participant number and model type) was given to ensure data was collected correctly. Upon setup completion, participants were asked to navigate to the app tutorial page to learn how to use their assigned model. The participants were then given an opportunity to use the app, including performing mock authentications and survey responses. Finally, the participants were asked to press the 'Begin the Study' button in the app, as well as schedule their exit meeting for one week later.

Upon completing this initial phase, participants were to go about their lives as usual, ensuring their 2FA tangible was with them and being aware of notifications from the study app. When a notification was received, participants were to open the app and authenticate with the tangible, with the option to skip if they did not have access to the tangible or for any other reason. After each authentication, the survey mentioned above was issued in the app to collect responses regarding the participant's experience using the 2FA tangible during that authentication.

Reflection Interview. Once the week had concluded, participants were met again, with this meeting taking place online. Thirty minutes were allocated for this meeting and proceeded as follows: The participants were reminded of the study's purpose and asked if they had any questions. A short survey was issued to collect demographic information. Then, a recorded interview was carried out to obtain information about the participant's experience using the tangible over the past week, their first impressions, non-assigned tangibles, and what they would like to see in the future for this method of authentication. Time was also allotted to allow participants to give any further information or ask final questions. For the

[2]To ensure an equal share of participants for each object, we had 15 objects in total, meaning that five participants interacted with each object.

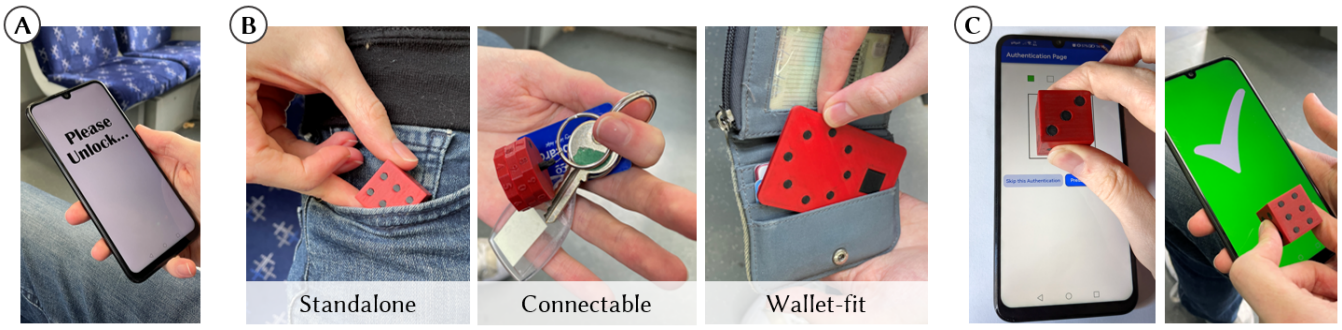


Figure 2: Tangibles used in Study II. To authenticate, users take out their smartphone (a), configure their tangible object (b) and then hold it against the touchscreen (c). For bigger depictions and screenshots of the app, we refer to Figure 3 in Appendix A.4.

full interview script, the reader is referred to Appendix A.2. Finally, the participants were reimbursed with an Amazon voucher with an equal value of roughly 25 US Dollars.

Pilot Study. Before the actual study, we conducted a week-long pilot study (N=3) to investigate the feasibility of the study design, as well as to detect any issues with the tangibles that impacted their performance over the course of extended use. Each participant was assigned one tangible. The tangibles were also intensively tested by the research team. The issues discovered during the study were related to the app developed and the tangibles created. One problem brought up was the lack of guidance and feedback offered by the app, to alleviate this an improved tutorial page was added, along with visual status hints for each tangible during authentication attempts. The cube model was also chosen to receive a progress indicator during the authentication attempts as this tangible required multiple touches to the screen.

Data Analysis. Before the analysis, all audio recordings were transcribed. Next, two researchers familiarised themselves with the data by reading the transcripts repeatedly. A shared codebook was proposed and finalised in a review meeting. The codebook is provided in Appendix A.3. Then, one author analysed all transcripts and applied the codebook reaching saturation after the 10th participant. After that, the second researcher further analysed the entire coding to validate and mark all the codings they disagreed with. Finally, the two researchers came together in a final discussion meeting to agree on a final coding. After this, both researchers grouped the codes into five main themes.

Recruitment, Participants. We recruited 15 participants through mailing lists and word-of-mouth. They were aged 26.1 years old (*min* = 21, *max* = 39, *median* = 24, *SD* = 5.43) on average. Ten identified as male, with the remaining five identifying as female. An affinity for technology interaction survey [13] was also issued, resulting in an average ATI score of 4.48 (*min* = 3.22, *max* = 5.89, *median* = 4.67, *SD* = 0.84). Hence, the sample had a rather high affinity for technology.

Limitations. We had a rather tech-savvy sample that might have overly welcomed the usage of tangibles. Consequently, our results can only serve as a first step towards understanding the design choices and interaction experiences of users. Hence, our results should be validated through future in-depth studies with more heterogeneous samples. The participants used tangibles randomly assigned to them. Hence, these tangibles were not personalisable. Because of that, participants might have received a tangible that they would not design themselves. Further, our tangibles were optimised for mobile usage considering the size. Because the size of a tangible might also be dependent on the device used, such as smartphones or tablets, future work should investigate the full process, including tangible design, personalisation, fabrication and usage as a whole including other sizes than those investigated by us. Further, participants used the tangible for one week. The results regarding ease of use and fun should be taken with a grain of salt since there might be a novelty effect. Consequently, future work should investigate longer usage periods. Finally, participants did not use the 2FA tangibles for their real accounts. This might have impacted their perceptions of the concept. Future work should investigate a realistic use case where 2FA tangibles protect real assets. However, our participants used the tangibles on their own devices and in different areas of their daily environments.

Ethical Considerations. All studies reported in this paper were reviewed and approved by our ethics board. The participants were informed via a consent form that participation is voluntary and that they could abort at any time without consequences. The collected data cannot be linked to individual participants. Audio data was transcribed before analysis, and the consent forms were kept separate from all other data. The recognition of tangibles would have been more accurate, and smaller tangibles would have been possible if par-

ticipants' mobile devices were rooted. This, however, would have exposed the private devices of the participants to security risks. In coordination with the ethics committee, we decided to simulate the security properties of 2FA tangibles with a low-resolution recognition to not expose the participants to security risks by rooting. Further, we used mock accounts for unlocking, because we did not want to impact the security of the participants' personal accounts. Further, the tangibles were printed with PLA plastic that, on the one hand, keeps its shape but, on the other hand, is soft enough to not injure participants or leave any kind of scratches on their devices.

6 Study II: Quantitative Results

Overall, the participants performed 197 authentications over the week and completed the related survey 163 times.

Authentication Success: The participants reported eleven unsuccessful authentication attempts. All of them were because the tangible was not available.

Duration: An authentication started once the participant indicated in-app that they have their tangible ready and ended when a complete attempt was recognised. For the cube, the mean duration of an authentication attempt was 15.5 seconds (*min* = 7.8, *max* = 31.7, *SD* = 5.38), for the card, it was 11.3 seconds (*min* = 5.4, *max* = 43.9, *SD* = 5.47) and for the pendant 8.6 seconds (*min* = 2.7, *max* = 44.8, *SD* = 8.09).

Location & Usage Intention: Participants were asked to indicate the location where they authenticated and whether they would like to perform an authentication in a similar setting to capture their usage intention. In total, 59 authentications (36.2%) were reported as *at home*. Of those, 64.4% answered affirmatively when asked about willingness to authenticate in a similar setting. For instance, P13 said they enjoyed the *"interactive way to authenticate items"*, while P4 liked that they could *"just have a place where [they] keep the authentication device"*. Five authentications (3.1%) were performed in *private places* different from their own home, with four participants stating usage intention. Eighty authentications (49.1%) were performed *at work* with 47.5% usage intentions. For example, P2 found that it *"takes too long"* to use for work, P5 also stated that it *"takes too much precision to operate"*. However, P4 felt that *"if [they] could keep the object just in the office ready, it would be handy"*. Six authentications (3.7%) were done *in transit* with 16.7% (N=1) usage intention. P8 stated that it was *"impractical whilst travelling"* to use the tangible, with P7 elaborating that they *"felt overwhelmed on the subway and couldn't concentrate on the activity"*. Finally, ten authentications (7.4%) were done in *public* with a 40% usage intention. P4 found they were *"not near... where [they] kept the [tangible]"* and found difficulty authenticating while P8 appreciated the aspect of *"security in a public place"*.

7 Study II: Qualitative Results

Overall, our participants welcomed the concept of 2FA tangibles for use in their daily lives. This section reports our results grouped into themes identified by the thematic analysis.

7.1 Security Perceptions

The participants liked that 2FA tangibles add a layer of security to their accounts. Metaphors also played an important role in defending from shoulder surfers or conveying security.

Security Benefits are Valued: Most of our participants commented on the security of 2FA tangibles, specifically considering that they are physically separate from the smartphone:

"This [tangible] combined with logging in to feels more secure. It does feel very secure. Like to have all the steps of having to have the thing and know the password for it and know which account it was linked to.", **P3**.

"For the most part I like the idea of the security being there. [gives examples] There is a bit of extra security because you've obviously got to have the token with you and then know how the token works. So, it's like an in-built two-factor authentication basically that requires you to have something and know how to use it.", **P8**.

The perceived benefit also impacted the participants' usage intentions. Most of them wanted to use 2FA tangibles for important accounts, such as financial services where 2FA is required and text messages with one-time passwords or apps were not considered secure enough. Further, participants frequently brought up the idea of using a 2FA tangible as a backup in case a password or other form of authentication mechanism was not available for them:

"[...] banking for example. Whenever I do, I tend to do financial things at home where I'm like, not in a hurry and I'm pretty stationary there, so in cases like that, for example, I would actually say yeah, why not? That could be good. A use case, I think, yeah.", **P12**.

Metaphors are Considered: Some participants linked their security perceptions to the specific tangible design that was considered as a metaphor. The cube was perceived as less secure than the card. The card, on the other hand, was perceived as less secure than the pendant. Participants stated that the pendant is already mentally associated with security and would prefer it based on the visual appearance. That was somewhat similar for the card shape. Here, some participants associated it with a credit card or they associated the interaction with the card with opening a safe by a security dial:

"I think the pendant [is my favourite] because it mimics already a lock. I think it makes you think that it's more secure than a dice where dice is almost just like a toy object.", **P4**.

The metaphors could also specifically be used to defend participants better. For instance, bystanders might associate cards and pendants with security. A few participants were concerned that bystanders might shoulder-surf them because they know they are currently authenticating. However, participants that used the cube did not voice such a concern because the cube is a neutral object. These comments also match some statements of participants in the online survey and related studies [17] who mentioned preferring benign everyday objects over something with a connection to security:

"I didn't really feel all that strange to be doing the authentication in public either like people probably just thought I was playing a game or something since it was just a red and black dice. I felt like it didn't really stand out much.", **P7**.

7.2 Positive Aspects

Our participants voiced further positive aspects after interacting with the tangibles for a week. In sum, the participants welcomed the possibility to customise the tangible and the option to self-fabricate it. The interaction was perceived as fun and easy to use. Finally, the tangible interactions were easy to memorise.

Customisation is Welcomed: Several participants particularly liked that 2FA tangibles are personalisable in a way. Hence, users can buy something different from a standardised off-the-shelf tangible. One participant even welcomed the independence from manufacturers since such tangibles could be 3D-printed by the users. Sample comments are:

"I like the basic concept of another factor for authentication that I can own. I'm a YubiKey user myself, so I guess I'm kinda well used to something like that and well. I thought about it and the idea of having a token that you can maybe customise even I think it's that's pretty cool for future ideas.", **P12**.

"I mean theoretically if I had issues with the model and it broke for something like a YubiKey, I'd need to go to the supplier and get a new one. But for something like this model, it's relatively easy to go off and print it for yourself and I feel that idea is really nice. It doesn't have any kind of really special technology that kind of limits it to not being able to be manufactured at home, and I feel that that's also a cool feature of it.", **P5**.

Using Tangibles is Fun: Several participants stated that authentication with the tangibles during the study period was fun for them. P3 gave a quite representative statement:

"I think it was quite successful. I just kept the card in my wallet and so I always had it with me when the thing went off. I enjoyed it generally. I just really enjoyed it, but it never really occurred to me in the past to like to have a physical thing that physical keys could be used for online accounts and so I like the idea that you can have this.", **P3**.

These results, however, should be taken with a grain of salt because the 2FA tangibles were only used for a week. Consequently, we cannot rule out potential novelty effects at this state of usage.

Tangibles are Easy-to-Use: Several participants considered the tangible and also the app as easy-to-use. They also considered that when we showed them the other tangibles that they had not used over the week. Some participants even commented on the tangibility:

"I felt quite good, I use MFA apps, so you know, where you use a PIN code, so you just copy and paste that. This is kind of the same idea. You basically just typing out a PIN, but a different method. So yeah, it was quite good.", **P13**.

"I like the idea that we don't have to even type something. This is the main advantage I think, in my opinion, that we don't type in anything. Any numbers, so.", **P11**.

Interactions are Memorable: Some participants said that the interaction with the tangibles was very easy to memorise for them compared to other traditional authentication methods:

"I think the pin and the system will be more memorable than a password for sure, if you had given me a password to log in for the week, I'd give you a month before I know what it was.", **P3**.

"Yeah, it's quite easy. I mean the PIN code was static, so it wasn't as if you have to remember like a randomised number each time, so just remember where it is in the dice and put in and you're finished.", **P12**.

7.3 Experienced and Perceived Issues

Even though the participants reported many positive aspects, they also told us about issues encountered during the week. These issues included forgetting the tangible at home, further design-related issues that might impact the interaction, problems with slippery tangibles, usability issues, and theoretical issues based on the fact that the 2FA tangibles are designed to work with touchscreen devices only.

Forgetting the 2FA Tangible: Some participants voiced issues based on the portability of 2FA tangibles. In particular, they were concerned about forgetting the tangible somewhere or the tangible being stolen and consequently being locked out of their accounts:

"Yeah, although it seems more secure, but still I have to carry this extra model. Uh, this is my concern. Probably I will. Sometimes I will. Forget to take it with me if it is integrated so I don't have to worry about whether sometimes probably, I will forget. So, just had to carry the model with me. So, if it is integrated then it should be very nice.", **P11**.

"I did not attach it to any of the things that I regularly take with me when I leave home. And because I had a week where I went to different places with different bags, I actually did not have it with me a lot of the time.", **P2**.

Issues with Shape or Size: Other participants had no issue with forgetting the tangible but voiced concerns based on the shape and size that it might impact portability. The cube, for instance, had too sharp edges for some participants, making them concerned about getting hurt while accidentally sitting on it in their pockets. While that did not happen during our study, the tangible's geometry was perceived to have a high impact on portability. One participant feared that sharp edges might even damage the smartphone screen, which is not possible due to the softness of the used PLA printing material:

"So, I generally like dice. [...] That's most of what I liked about it. Uhm. I think it's not that handy because you can't really transport it. Uh, because it has very sharp edges and so on, and it's pretty big for a device.", **P9**.

"So the first thing my partner commented on when I took the thing home and tried to authenticate at home was like, yeah, but you have to touch it onto your phone and if you like if it is dirty because it's coming out of your bag, what if you are leaving scratches in your phone display?", **P2**.

Further, the size of the tangible was frequently mentioned. E.g., the cube was too big to match the habit of not carrying anything besides the smartphone of P7. Another example was P9, who did many interactions with their smartwatch, but the cube was too big for that. The concerns voiced by these participants were not linked to the concept of 2FA tangibles in general but rather to the specific dimensions that tangibles could or should have:

"So because of the size of the object, the display of the Smartwatch is too tiny for that kind of authentication. So doing it with a smartphone as the smartphone is the main device when accessing services, I would say yes. No, I don't want to use it on my smartwatch.", **P9**.

Slippery Tangibles: Some participants reported issues using the tangible with one hand on the go because the tangible slowly slid away. P10 gave a representative comment:

"When I used it at home, it would be much easier for me than to use it outside, because you need to have the phone placed at the table or any non-moving object and you need to press down the authentication object on it. So, for example, if I was at, uh, in a bus or on any uh, movable. Uh, sorry, in any uncontrolled situation, I don't think the authentication method will have worked.", **P10**.

Interoperability Aspects: An issue that was not actively experienced by the participants but was frequently mentioned in the interviews was the interoperability of 2FA tangibles with other devices. Since the tangibles investigated by us are limited to touchscreen devices, some participants wished for better interoperability, including laptops or personal computers without touchscreen functionality.

"Generally for interacting with a computer, I don't feel it's the best for interaction with, a non-touchscreen device just due to how it's implemented. Theoretically, if it also worked on the trackpads as well, perhaps it could be used for authentication with laptops.", **P5**.

Usability Issues: Moreover, the participants also reported usability issues that were mostly linked to the way the tangibles need to be used. These participants mainly had issues when using the tangibles on the go:

"From a user perspective, it's very inconvenient [...] it's not something which is easy, you can't do it while holding the telephone free and I've had my best success rate when fixating or putting the phone on a desk something and setting the authentication object on it and which basically means you have to sit somewhere [...] And if you're not in a situation which allows this, some kind of setting it's just difficult.", **P1**.

"Usability wise [it] is like, if you're just sitting nice, it's nice and easy to get out if you're not moving and as I say it then becomes a factor of the location.", **P8**.

8 Discussion & Limitations

This section first discusses the security properties of 2FA tangibles including a path to realistic tangibles that are secure. Next, we focus on the portability issues voiced by the participants and options to solve them. This is followed by a discussion of the tangible shapes, sizes as well as considerations thereof and limitations of our investigations. Finally, we use this discussion to motivate a user-centred fabrication pipeline for 2FA tangibles.

What About Security? First and foremost, security is the most important aspect of authentication [3]. While the tangibles used in our investigation can only provide limited security, as their authentication patterns are quite simple, it was sufficient for investigating how 2FA tangibles integrate into daily life from an HCI perspective. To create secure items, the following challenges need to be solved:

Completely 3D-printed tangibles: We 2FA tangibles should be completely 3D-printed, we need a way to *increase* the resolution of the authentication pattern to offer a large password space by encoding a more significant number of different interactions. This requires access to the capacitive raw data [27, 28] which currently is not possible on non-rooted devices. The tangibles used in our study were limited to the Android API that only offers access to ten touchpoints – one for each finger – at once. Having access to this data allows the recognition of a larger number of smaller dots in closer distances to each other. Hence, device manufacturers need to provide more powerful APIs that give more options to developers. While this challenge may be technologically solvable in the long run with device manufacturer support, the question arises whether fully 3D-printed tangibles are indeed an ideal solution for realising 2FA. Having fully 3D-printed tangibles has several benefits: (1) tangibles can be printed in one pass

without the need to configure any electronics, (2) the tangibles are passive, removing the need for energy sources, and (3) shapes can be personalised. The second two aspects were also mentioned by our study participants, with one person even liking the idea of 3D printing the tangible at home because it provides independence from suppliers. Since the current recognition technology is not yet accurate enough, we would also like to argue for partially 3D-printed tangibles.

Partially 3D-printed tangibles: The personalisation benefits from 3D-printing could be combined with other technology, e.g., NFC tags or passive tokens [35]. The general idea is that tangible interactions are used to activate the token. Currently, YubiKey tokens need a simple touch making the possession of the token sufficient to impersonate the user. To address this issue, we envision a series of more complex interactions to trigger authentications. Even though this has to be verified by future work, such tangibles would likely have similar security perceptions as those in our study. Further, they would solve the interoperability issues voiced by several participants since such tags are not limited to touchscreens.

In summary, the security of 2FA has priority but has yet to be realised by standalone 3D printing. Therefore, we recommend combining the usability and UX benefits of 3D-printed 2FA tangibles with the security benefits of other technology, such as NFC tags.

Addressing Portability Issues. Many participants voiced concerns about the portability of 2FA tangibles. Some participants even forgot the tangibles at home and could not access them for some time during our study. These concerns were mainly linked to standalone tangibles (the cube in Study II) that can neither be connected to another object nor fit into a wallet or pocket. Moreover, several participants stated they were unwilling to do security-critical interactions, like banking transactions, on the go unless it is urgent. Consequently, they would not need tangibles with great portability. Based on that, we recommend integrating the location of intended tangible use in the design pipeline, such that portability can be considered. Portable tangibles should either be small enough to easily fit in a pocket or wallet or offer an option to be connected to another object, like a key chain. Shape properties should be considered, such that users do not get injured from too sharp edges while having a tangible in the pocket.

Another interesting aspect is that users are not limited to a specific material that works in any environment. As suggested by one of our participants, users might have a static tangible on their desk or at another place at home for most interactions and a mobile tangible that they keep in their wallet in case they have to authenticate on the go. Since this contrasts study results in the literature where study participants did not want multiple items [31], future work has to validate this. However, related work specifically investigated one-time password generators with a specific form factor and size. Hence, these results might not transfer to 2FA tangibles.

Tangible Shapes and Sizes. The majority of the over 200 participants in our first study chose simple geometric shapes. Animals, like cats or dogs, formed the most complex shapes. The participants of the main study revealed more in-depth considerations of that. Tangibles with sharp edges might hurt users when they sit on them or look for them in their pockets. Further, complex shapes might be more likely to break and do not fit well into pockets or wallets. The sizes of tangibles were also closely connected to portability. Highly portable tangibles should be small, but those used at home could be bigger, with some participants in the online study even designing tangibles of 10 cm size that could cover their entire smartphone screen. Based on that, we conclude that the shape of 2FA tangibles should be simple. This does not mean that simple geometric shapes are the only solution; animals and other shapes that people like could also be simplified. As for the size, again, the environment in which the user intends to use the tangible should be carefully considered.

What About the Friction? Friction [20] is a construct from habit research denoting anything that might constrain a human in doing a specific task. Friction might be beneficial to get rid of unhealthy habits but might be an obstacle to creating new ones. Throughout the reports of the participants, we found two ways how 2FA tangibles impact the participants' habits. Either the tangibles resulted in more friction because they did not match user habits, or they helped them establish new and helpful security habits that they have not had before.

We further investigated the usage over the course of a week to also find out whether participants could establish a habit of using the tangibles and how the tangibles might interfere with other habits. As detailed in the results section, only two participants struggled with bringing the tangible with them because it was their habit to carry only their smartphones and wallets. Both participants interacted with the cube that neither fits in a wallet nor can it be attached to another object. When confronted with the other models during the interviews, the participants were more positive, yet honestly stated that their habits would likely prevent them from using external devices for authentication for two reasons: First, participants stated to prefer performing security critical tasks in a static environment, for instance, at home. Second, they were used to purely digital solutions, e.g., one-time passwords by text messages, that they would need more time to create the habit of interacting with 2FA tangibles. The remainder of the participants struggled less. They either had the tangibles all the time in their pockets, on their key rings, in their wallets or placed them in a dedicated spot, for instance, on their desk to, be available for authentication. They were used to carrying more when they left home, so the additional tangible did not add more friction. Two participants even went on vacation during our study and brought the tangibles with them without issues. Since the tangibles better matched the habits of these participants, they had fewer issues in performing the study

tasks on the go and even welcomed this new security habit.

In sum, if users have the habit of not carrying any specific objects besides their smartphones regularly, then 2FA tangibles create more friction. For users that already carry additional objects, like a purse or key rings, 2FA tangibles are connected to an existing habit and are available when needed.

9 User-Centred Fabrication Pipeline

Based on our results and the discussion, we envision a user-centred fabrication pipeline where the users design personalised 2FA tangibles. Per step, we highlight either possible realisations or provide guidance for future work.

1) Usage Type. First, the users choose whether they want to use their 2FA tangible in a static or mobile fashion because this has implications for the tangible design, shape, and size. This step does not require further investigations, however, future work should investigate what users primarily choose as environment type.

2) Usage Context. Next, users indicate the specific surrounding environment. E.g., static environments might be their home which has other implications for security compared to a shared workspace. Here, a list of possible specific environments is required. Our study participants stated to use the tangibles at home, at work, on the go, or in other private environments. Further, security implications of these specific environments are needed. At home, for instance, shoulder surfing might be less of an issue. These security implications should be the basis for an environment-specific security model that helps users choose suitable interactions later on.

3) Device. The device the user wants to use the tangible for impacts the tangible's shape and further properties. For instance, if the tangible needs to be recognised on a touchscreen, the tangible requires at least one flat surface. If the tangible would be used with a PC, it might need a USB connector. While the requirements for standalone 3D-printed devices and USB keys are partly known, future work should investigate the requirements for mixed tangibles that allow a custom shape with integrated sensors.

4) Interactive Choice of Tangible Shape & Size. Since the shape impacts the tangible size, users should be able to choose these two properties at the same time interactively. The pipeline should have a list of suggested shapes based on the previous two choices. For static environments, a standalone tangible might be the first suggestion, whereas wallet-fit or connectable tangibles might be better for mobile users.

5) Interaction(s). Once the shape and size are known, users suggested that they want to first decide on the number of authentication interactions. For this, the pipeline should give a recommendation based on the security properties of the specific environment. Using this information, the pipeline can automatically calculate possible interactions that can be performed with the chosen shape. Especially the last step requires an algorithm that takes a 3D shape as input and calculates possible interactions based on it, offering another opportunity for future work.

6) Manufacturing. The users now choose to fabricate the tangible themselves, go to a public maker space or order it from a manufacturer. This is based on the specific components required to fabricate a tangible.

7) Backup. Finally, all design decisions are stored in an encrypted file to allow revoking and recreating the 2FA tangible in case it was lost or broken.

Exploration of this fabrication pipeline and its evaluation with users is a mission for future research toward user-friendly tangible 2FA that can be customised to the user and use case.

10 Conclusion

2FA tangibles are a potentially viable alternative for solving UX and security issues of currently available 2FA mechanisms. To investigate 2FA tangibles, we first simulated a simple fabrication pipeline where 226 participants designed tangibles by describing their size, colour, shape, and possible interaction. Participants' designs mainly consisted of simple geometric shapes that either described a) standalone objects, b) tangibles that can be connected to another object, or c) tangibles that fit into wallets or pockets.

For each of those categories, we prototyped one tangible and let 15 participants use our tangibles in the wild to perform authentications over one week. From our study, we learned that the participants welcome the security benefit provided by the 2FA tangibles. Further, they considered the specific tangible design as a metaphor that could support security perceptions or obscure the connection to security. The main issues that participants experienced during the study were connected to portability, but those participants would prefer using a tangible in a static environment, such as their desk at home. Based on the results of our investigations, we first discussed possibilities to address the shortcomings and how 2FA tangibles impact user habits. Finally, we proposed a user-centred fabrication pipeline that can be used by the users to design personalisable 2FA tangibles. Overall, 2FA tangibles are a promising solution to make 2FA easier to use, fun and more secure, but future work is needed to realise fabrication pipelines and investigate them with users.

Acknowledgments

This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC 2092 CASA - 390781972. Furthermore this work was co-funded by the EPSRC(EP/V008870/1).

References

[1] Jacob Abbott and Sameer Patil. How mandatory second factor affects the authentication user experience. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–13, New York, NY, USA, 2020. Association for Computing Machinery.

[2] Claudia Ziegler Acemyan, Philip Kortum, Jeffrey Xiong, and Dan S Wallach. 2fa might be secure, but it's not usable: A summative usability assessment of google's two-factor authentication (2fa) methods. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 62(1):1141–1145, 2018.

[3] Joseph Bonneau, Cormac Herley, Paul C. Van Oorschot, and Frank Stajano. The quest to replace passwords: A framework for comparative evaluation of web authentication schemes. In *Proceedings of the IEEE Symposium on Security and Privacy*, pages 553–567, Piscataway, NJ, USA, 2012. IEEE.

[4] Swati Chaudhari, SS Tomar, and Anil Rawat. Design, implementation and analysis of multi layer, multi factor authentication (mfa) setup for webmail access in multi trust networks. In *Proceedings of the International Conference on Emerging Trends in Networks and Computer Communications (ETNCC)*, pages 27–32, Piscataway, NJ, USA, 2011. IEEE.

[5] Stéphane Ciolino, Simon Parkin, and Paul Dunphy. Of two minds about two-factor: Understanding everyday FIDO u2f usability through device comparison and experience sampling. In *Proceedings of the Symposium on Usable Privacy and Security (SOUPS)*, pages 339–356, Berkeley, CA, US, August 2019. USENIX Association.

[6] Jessica Colnago, Summer Devlin, Maggie Oates, Chelse Swoopes, Lujo Bauer, Lorrie Cranor, and Nicolas Christin. "it's not actually that horrible": Exploring adoption of two-factor authentication at a university. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 456:1–456:11, New York, NY, USA, 2018. ACM.

[7] Federal Financial Institutions Examination Council. Authentication in an internet banking environment. *Retrieved June*, 28:2006, 2005.

[8] Alexei Czeskis and Juan Lang. Fido nfc protocol specification v1.0. FIDO Alliance Proposed Standard, 2015.

[9] Sanchari Das, Andrew Dingman, and L. Jean Camp. Why johnny doesn't use two factor a two-phase usability study of the fido u2f security key. In *Proceedings of the International Conference on Financial Cryptography and Data Security (FC)*, pages 1–20, Cham, Switzerland, 2018. Springer.

[10] Sanchari Das, Gianpaolo Russo, Andrew C. Dingman, Jayati Dev, Olivia Kenny, and L. Jean Camp. A qualitative study on usability and acceptability of yubico security key. In *Proceedings of the 7th Workshop on Socio-Technical Aspects in Security and Trust*, STAST '17, page 28–39, New York, NY, USA, 2018. Association for Computing Machinery.

[11] Emiliano De Cristofaro, Honglu Du, Julien Freudiger, and Greg Norcie. A comparative usability study of two-factor authentication. In *Proceedings of the Workshop on Usable Security*, pages 1–10, 2014.

[12] J. Dutson, D. Allen, D. Eggett, and K. Seamons. Don't punish all of us: Measuring user attitudes about two-factor authentication. In *Proceedings of IEEE European Symposium on Security and Privacy Workshops (EuroS PW)*, pages 119–128, Piscataway, NJ, USA, 2019. IEEE.

[13] Thomas Franke, Christiane Attig, and Daniel Wessel. A personal resource for technology interaction: Development and validation of the affinity for technology interaction (ati) scale. *International Journal of Human-Computer Interaction*, 35(6):456–467, 2019.

[14] S. Ghorbani Lyastani, M. Schilling, M. Neumayr, M. Backes, and S. Bugiel. Is fido2 the kingslayer of user authentication? a comparative usability study of fido2 passwordless authentication. In *Proceeedings of the IEEE Symposium on Security and Privacy (SP)*, pages 268–285, Piscataway, NJ, USA, 2020. IEEE.

[15] Maximilian Golla, Grant Ho, Marika Lohmus, Monica Pulluri, and Elissa M Redmiles. Driving 2fa adoption at scale: Optimizing two-factor authentication notification design patterns. In *Proceedings of the USENIX Security Symposium*, USENIX Security 21, pages 109–126, Berkeley, CA, US, 2021. USENIX Association.

[16] Kat Krol, Eleni Philippou, Emiliano De Cristofaro, and M Angela Sasse. "they brought in the horrible key ring thing!" analysing the usability of two-factor authentication in uk online banking. In *Proceedings of the Workshop on Usable Security*, USEC 2015, Reston, VA, USA, 2015. Internet Society.

[17] Karola Marky, Kirill Ragozin, George Chernyshov, Andrii Matviienko, Martin Schmitz, Max Mühlhäuser, Chloe Eghtebas, and Kai Kunze. "nah, it's just annoying!" a deep dive into user perceptions of two-factor authentication. *ACM Trans. Comput.-Hum. Interact.*, 29(5), oct 2022.

[18] Karola Marky, Martin Schmitz, Verena Zimmermann, Martin Herbers, Kai Kunze, and Max Mühlhäuser. 3d-auth: Two-factor authentication with personalized 3d-printed items. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–12, New York, NY, USA, 2020. Association for Computing Machinery.

[19] Philipp Mayring. *Qualitative content analysis: theoretical foundation, basic procedures and software solution*. Social Science Open Access Repository (SSOAR), Klagenfurt, 2014.

[20] Asaf Mazar, Geoffrey Tomaino, Ziv Carmon, and Wendy Wood. Sustaining sustainability: Lessons from the psychology of habits. *PsyArXiv Prepr*, 2020.

[21] Martez Mott, Thomas Donahue, G. Michael Poor, and Laura Leventhal. Leveraging motor learning for a tangible password system. In *Extended Abstracts of the CHI conference on Human Factors in Computing Systems*, CHI EA '12, pages 2597–2602, New York, NY, USA, 2012. ACM.

[22] Martez Mott, Thomas Donahue, G. Michael Poor, and Laura Leventhal. Leveraging motor learning for a tangible password system. In *CHI '12 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '12, page 2597–2602, New York, NY, USA, 2012. Association for Computing Machinery.

[23] Ahmad R. Pratama and Firman M. Firmansyah. Until you have something to lose! loss aversion and two-factor authentication adoption. *Applied Computing and Informatics*, 2021.

[24] Ken Reese, Trevor Smith, Jonathan Dutson, Jonathan Armknecht, Jacob Cameron, and Kent Seamons. A usability study of five two-factor authentication methods. In *Fifteenth Symposium on Usable Privacy and Security*, SOUPS 2019, Berkeley, CA, US, 2019. USENIX Association.

[25] Jun Rekimoto. SmartSkin: An Infrastructure for Freehand Manipulation on Interactive Surfaces. In *Proc. SIGCHI Conference on Human Factors in Computing Systems*, CHI '02, pages 113–120. ACM, 2002.

[26] Joshua Reynolds, Trevor Smith, Ken Reese, Luke Dickinson, Scott Ruoti, and Kent Seamons. A tale of two studies: The best and worst of yubikey usability. In *Proceedings of the IEEE Symposium on Security and Privacy (SP)*, pages 872–888, Piscataway, NJ, USA, 2018. IEEE.

[27] Martin Schmitz, Florian Müller, Max Mühlhäuser, Jan Riemann, and Huy Viet Viet Le. Itsy-bits: Fabrication and recognition of 3d-printed tangibles with small footprints on capacitive touchscreens. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery.

[28] Martin Schmitz, Jürgen Steimle, Jochen Huber, Niloofar Dezfuli, and Max Mühlhäuser. Flexibles: Deformation-Aware 3D-Printed Tangibles for Capacitive Touchscreens. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '17, pages 1001–1014, New York, NY, USA, 2017. ACM.

[29] DUO Security. Security tokens. https://duo.com/product/trusted-users/two-factor-authentication/authentication-methods/security-tokens, 2019. [Online; accessed: 22-August-2019].

[30] Teddy Seyed, Xing-Dong Yang, Anthony Tang, Saul Greenberg, Jiawei Gu, Bin Zhu, and Xiang Cao. Cipherchard: A token-based approach against camera-based shoulder surfing attacks on common touchscreen devices. In *Human-Computer Interaction – INTERACT 2015*, page 436–454, Berlin, Heidelberg, 2022. Springer-Verlag.

[31] Jake Weidman and Jens Grossklags. I like it, but i hate it: Employee perceptions towards an institutional transition to byod second-factor authentication. In *Proceedings of the Annual Computer Security Applications Conference*, ACSAC 2017, pages 212–224, New York, NY, USA, 2017. ACM.

[32] Catherine S. Weir, Gary Douglas, Martin Carruthers, and Mervyn Jack. User perceptions of security, convenience and usability for ebanking authentication tokens. *Computers & Security*, 28(1-2):47–62, 2009.

[33] Catherine S. Weir, Gary Douglas, Tim Richardson, and Mervyn Jack. Usable security: User preferences for authentication methods in ebanking and the effects of experience. *Interacting with Computers*, 22(3):153–164, 2009.

[34] Shota Yamanaka, Kunihiro Kato, Tung D Ta, Kota Tsubouchi, Fuminori Okuya, Kenji Tsushio, and Yoshihiro Kawahara. Sheetkey: Generating touch events by a pattern printed with conductive ink for user authentication. 2020.

[35] Yubico. Yubikey neo. https://www.yubico.com/products/yubikey-5-overview/, 2019. [Online; accessed: 25-August-2022].

A Material Study II

A.1 In-App Questions

When login was successful:

- Where are you at the moment?
  - at home
  - at work
  - in transit (e.g., bus or train)
  - in a public place (e.g., restaurant, park)
  - in a private place (e.g., home of a friend)
  - other (please specify)
- Did you experience any issues with authentication?
  - yes
  - no
- What kind of issues did you experience?
  - I had to look for the item
  - I needed multiple attempts
  - The timing of authentication was not convenient
  - Other (please specify)
- Would you like to perform 3D authentication in a similar setting in your daily life?
  - yes –> why?
  - no –> why not?
- Do you have any additional feedback? (free text)

When login was not successful:

- Where are you at the moment?
  - at home
  - at work
  - in transit (e.g., bus or train)
  - in a public place (e.g., restaurant, park)
  - in a private place (e.g., home of a friend)
  - other (please specify)
- What kind of issues did you experience?
  - I had to look for the item
  - I needed multiple attempts
  - The timing of authentication was not convenient
  - Other (please specify)
- Would you like to perform 3D authentication in a similar setting in your daily life?
  - yes –> why?
  - no –> why not?
- Do you have any additional feedback? (free text)

When login had to be skipped (when pressing the skip login button):

- Where are you at the moment?
  - at home
  - at work
  - in transit (e.g., bus or train)
  - in a public place (e.g., restaurant, park)
  - in a private place (e.g., home of a friend)
  - other (please specify)
- What kind of issues did you experience?
  - I had to look for the item
  - I needed multiple attempts
  - The timing of authentication was not convenient
  - I accidentally skipped
  - Other (please specify)
- Would you like to perform 3D authentication in a similar setting in your daily life?
  - yes –> why?
  - no –> why not?
- Do you have any additional feedback? (free text)

A.2 Interview Script

Thanks for participating in our study. During the past week, you have used one of the 3D-printed authentication items. In this interview, we would like to learn about our experience to improve the items to create better authentication mechanisms in the future. We are interested in your opinion, there are no right or wrong answers.

- How was the last week when you interacted with the items?
- Were there any issues? If yes, which ones? (Talk about each issue and ask what could be improved, separate between app and item)
- What did you like about the items?
- What didn't you like? How could that be made better in your opinion?
- Here are two alternatives that we designed, have a look at them. Compared to your item, would you prefer one of the alternatives? Why (not)?
- The 3D printed items can be printed in any shape, if you could decide, what would you prefer and why?
- Assuming that your dream item would be possible, would you like to use it in your daily life or rather something else like SMS notifications or a USB Key? Why (not)?
- Is there anything else that you would like to tell us?

A.3 Codebook

Table 1: Codebook used to analyse the interviews.

Category	Code	Description
Security Perceptions	separated_token	security benefit by a separated token
	metaphor	security consideration of a metaphor
	eyes_free_interaction	tangibles might be used in secret without looking at it
	observing	considerations based on presence of bystanders
	scalability	secures many devices
Positive Aspects	customisation	tangibles can be customised or self-fabricated
	interaction_fun	tangibles are fun to use
	ease_of_use	tangibles are easy to use
	memorability	interactions are easy to remember
Experienced Problems	tangible_forgotten	tangible was forgotten, e.g., at home
	tangible_size	problems based on size
	tangible_design	problems based on shape
	no_recognition	tangible was not recognised by app
	usability	Manufacturer is responsible
Considered Issues	tangible_might_be_forgotten	tangible might be forgotten
	reset_needed	tangible needs reset after login
	tangible_too_big	tangible perceived too big
	tangible_design	design not liked
	interoperability	tangible limited to touchscreen device
Ideal Tangible	thin	tangible should be thin
	small	tangible should be small
	everyday_item	tangible should be everyday item
	durable	tangible should be durable
	connectable	tangible should be connectable

A.4 Prototypes and Screenshots

In this section, we provide screenshots of the app used in Study II and pictures of the 2FA tangibles.

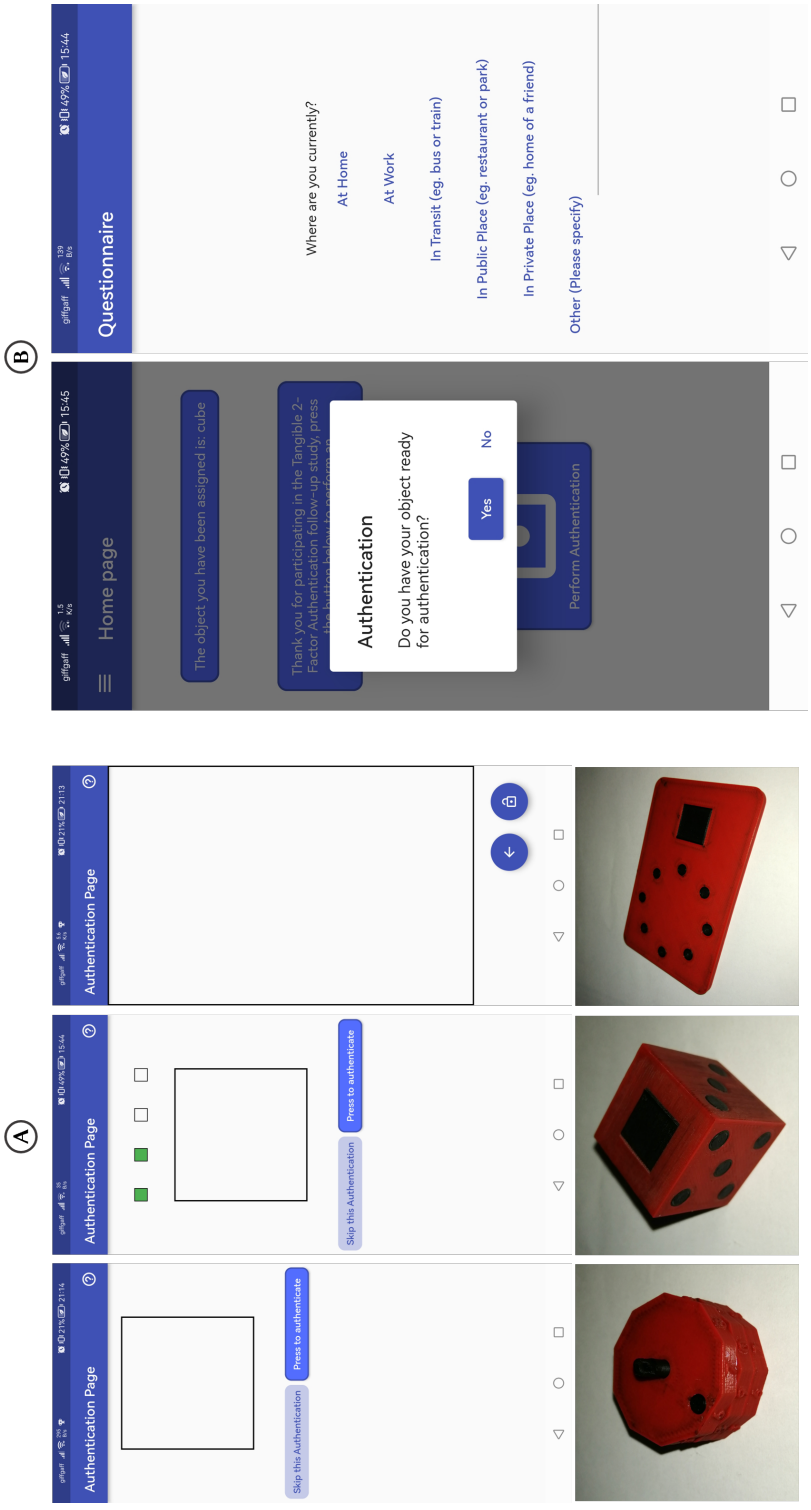


Figure 3: This figure depicts our three developed prototypes as well as screenshots of the app used in the study. Part A shows the specific login screen for each tangible whereas part B shows the screen of the survey after the authentication.

Prospects for Improving Password Selection

Joram Amador
Pomona College

Yiran Ma
Pomona College

Summer Hasama
Pomona College

Eshaan Lumba
Pomona College

Gloria Lee
Pomona College

Eleanor Birrell
Pomona College

Abstract

User-chosen passwords remain essential to online security, and yet users continue to choose weak, insecure passwords. In this work, we investigate whether *prospect theory*, a behavioral model of how people evaluate risk, can provide insights into how users choose passwords and whether it can motivate new designs for password selection mechanisms that will nudge users to select stronger passwords. We run a pair of online user studies, and we find that an intervention guided by prospect theory—which leverages the reference-dependence effect by framing a choice of a weak password as a loss relative to choosing a stronger password—causes approximately 25% of users to improve the strength of their password (significantly more than alternative interventions) and improves the strength of passwords users select. We also evaluate the relation between feedback provided and password decisions and between users' mental models and password decisions. These results provide guidance for designing and implementing password selection interfaces that will significantly improve the strength of user-chosen passwords, thereby leveraging insights from prospect theory to improve the security of systems that use password-based authentication.

1 Introduction

User-chosen passwords remain a critical component of security. Many efforts have been made to nudge users towards choosing stronger passwords, including password rules [33] and password meters [20], but these efforts have met with only partial success. Password rules are ineffective at enforcing strong password choices [33, 69], many password meters are ineffective [13] especially for accounts users consider unimportant [20], and users continue to select and use weak passwords [45]. In this work, we investigate the extent to which insights from behavioral economics apply to users' password selection decisions and how those insights might be leveraged to enhance security by nudging users to select stronger passwords.

Prospect theory [32, 60–63] is an empirically-grounded behavioral model of how people make decisions in the presence of risk. Prospect theory has been applied to various different areas of economics; it has proven a successful model both for explaining observed behaviors [8, 12, 16, 28, 37, 38, 44, 54, 55] and for prescriptively nudging people towards higher-utility choices [24, 29, 39, 59].

Interactions between humans and systems that affect security and privacy can be framed as decisions in the presence of risk. For example, password selection requires users to evaluate the risk associated with each possible password they consider (how likely is it that their account will be compromised if they select that password and how bad will the consequences be if that occurs) and balance that risk against other competing factors (e.g., memorability and easy of typing, including on mobile devices) in order to decide which password to use. However, prior work has thus far explored the intersection between prospect theory and security and privacy only in limited specific domains, such as investment in security [53, 66], adoption of two-factor authentication [48], disclosure of personal information [3, 4, 27], cookie consent [42], and tracking authorization [18]. In this work, we explore the connection between prospect theory and password selection through a pair of online user studies.

Our first study explores the connection between two effects identified in the prospect theory literature—the *reference-dependence effect* and the *source-dependence effect*—and password selection. We ran an online user study on Amazon Mechanical Turk with 762 participants in which we asked people to create an account on an experimental website. Users who initially selected weak passwords or moderate passwords

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.
USENIX Symposium on Usable Privacy and Security (SOUPS) 2023.
August 6–8, 2023, Anaheim, CA, United States.

were presented with an interactive prompt asking whether they wanted to go back and choose a stronger password; there were six different versions of the interactive prompt corresponding to three different framings (positive, neutral, and negative) and two different prompt phrasings (specific and vague). Participants also completed a follow-up survey about their beliefs regarding passwords and password-related risks. We found that the reference-dependence effect applies to password selection decisions—i.e., an interaction with negative framing resulted in significantly higher rates of improvement compared to neutral framing ($p < .001$) or positive framing ($p = .027$). However, the source-dependence effect did not appear to apply; the phrasing of the prompt (specific of vague) did not have a significant impact on whether user went back and selected a stronger password.

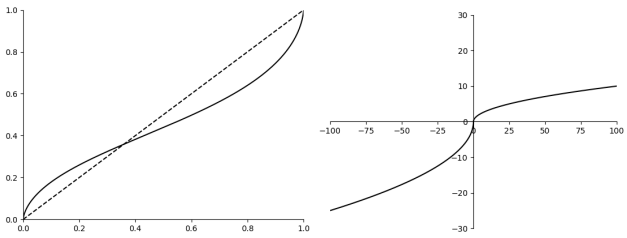
To validate the reference-dependence effect and to further understand how it influences password selection, we conducted a second user study through Prolific ($n = 607$) in which we recorded fine-grained measurements about password strength—as measured by the `zxcvbn` password meter's estimate of the number of guesses it would take to crack each password—along with information about how people modified their passwords. We also explored the impact of feedback and suggestions by including a condition with no meter and conditions in which the interactive prompt included suggestions for improving the password. Our results validated the reference-dependence effect for password selection decisions and provided insight into how people change their passwords in response to such interventions. We did not observe any significant differences due to feedback or suggestions.

Finally, we investigated whether mental models of security affected how users responded to our interactions. We found that perceptions about likely targets are correlated with password selection decisions but that decisions were consistent across different models of risks.

Our results suggest that some prospect theory effects can provide a model for understanding users' password selection decisions. In particular, we found that an intervention that leverages negative framing can significantly strengthen passwords. We believe that this insight from prospect theory can form the foundation for designing and implementing password selection mechanisms that enhance security by nudging users to select stronger passwords.

2 Background: Prospect Theory

Prospect theory [32,60–63]—first introduced in the 1970s as a critique of the then-dominant expected utility theory [23,67]—is a descriptive model of decision making in the presence of risk. Expected utility theory—which asserts that a principal faced with a choice between two options will evaluate the expected utility of each outcome and then select the option with the higher expected utility—does not accurately predict human behavior observed in many experimental settings.



(a) Decision weight function w (b) Subjective value function ν

Figure 1: Example functions matching empirically-observed behavior proposed by prior work [7, 63].

Prospect theory instead posits that decisions are comprised of two phases: an editing phase and an evaluation phase. In the editing phase, humans apply a set of simplifying heuristics to reduce the complexity of the decision problem. In the evaluation phase, probabilities and utilities are weighted by a decision weight w and a subjective value ν, respectively; example functions capturing empirically-observed behavior are shown in Figure 1. Humans are then presumed to rationally evaluate the options based on the weighted expected subjective value of the edited prospects.

The interactions between the editing phase and the weighting functions w and ν result in several effects that have been empirically validated through a series of experimental studies:

1. *Isolation Effect:* People simplify decision problems by disregarding components shared between alternatives and focusing exclusively on components that distinguish the options.

2. *Pseudocertainty Effect:* People simplify decision problems by treating extremely likely (but uncertain) outcomes as though they were certain.

3. *Reference-dependence Effect:* People simplify decision problems by defining outcomes relative to a neutral baseline. The framing of a problem can effect which baseline is used.

4. *Certainty Effect:* People overweight the probability of outcomes that are certain relative to outcomes that are merely probable.

5. *Source-dependence Effect:* People have different decision weights depending on the type of risk. For example, people have higher decision weights for contingent risks than for equivalent probabilistic risks (e.g., they prefer an insurance policy that provides certain coverage of specific types of damages to one that provides probabilistic coverage of all types of damages). Similarly, people are *ambiguity averse*—they prefer to bet based on precisely defined odds rather than on unknown odds.

6. *Loss Aversion Effect:* People subjectively dislike losses more than they value gains. That is, the value function is steeper for negative values (losses) than for positive values (gains).

More than 40 years later, prospect theory is still widely viewed as the best available model for how people make decisions in the presence of risk. It has been applied as a descriptive model to explain observed behavior in various different areas of economics including finance [6, 19, 44, 54], insurance [8, 30, 37, 56], savings [38], price setting [28], labor supply [12, 16], and betting markets [55]. Within the domain of computer science, prospect theory has been applied to explain decisions relating to investment in security [53, 66], adoption of two-factor authentication [48], disclosure of personal information [3, 4, 27], cookie consent [42], and tracking authorization [18].

Prospect theory has also been applied prescriptively in certain domains to nudge people towards certain "desirable" behaviors, including nudging employees to increase their retirement contributions [59], encouraging teachers to improve student outcomes [24], and incentivizing teams in high-tech factories to increase their productivity [39]. However, prospect-driven interventions have not been uniformly successful: a 2012 study did not see any increase in effort when financial or non-financial incentives for students were framed as losses compared to equivalent incentives framed as gains [29], and a 2021 study found that framing did not significant effect user decisions about whether to authorize tracking by iOS apps.

3 Related Work

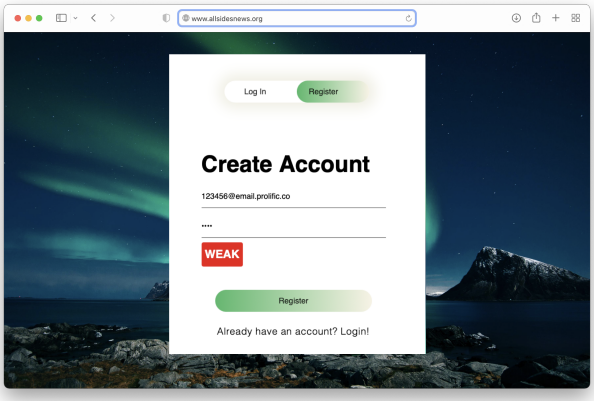
Improving Password Selections. Given the prevalence of password-based authentication and the ongoing dependence on user-chosen passwords, a large body of work has been dedicated to improving the strength of passwords that users select.

Early work on estimating password strength generally focused on entropy-based metrics [36]. However, entropy has since been criticized as been a poor measure of password guessability [33, 69, 70]. More recent efforts use dictionaries of words, lists of leaked passwords, and variants of words in those dictionaries and lists (e.g., L33t-style substitutions or addition of common suffixes) to define classes of weak or prohibited passwords [25, 43, 70].

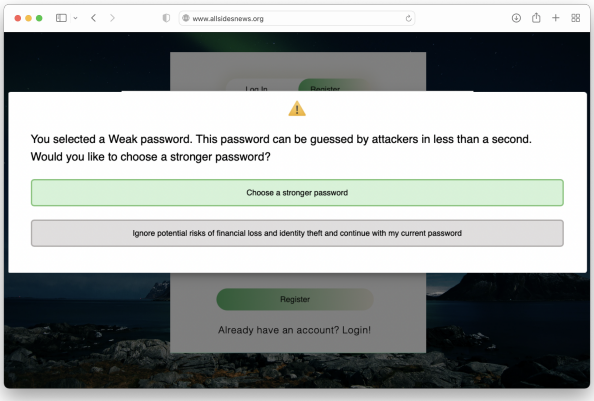
Studies have found that users exhibit misconceptions about password strength [65], which has resulted in increasing adoption of password meters across the most popular websites [20]. In general, having a password meter improves password strength, especially for accounts that users consider important [20]. However, some websites continue to use metrics that rely on entropy-based metrics and are thus inconsistent at effecting strong password selections [69]; one study found that most password meters deployed on actual websites are ineffective [13]. Careful calibration is also required to ensure that usability considerations do not undermine the benefits of a password meter: meters that are too strict can annoy users, while meters that are too lenient can result in weaker password selections [64].

Applications of Prospect Theory to Security. Despite the success of prospect theory in economics, there has been limited work applying prospect theory to security decisions, and only in limited domains. Verendel [66] developed a prospect theory model for decisions about buying versus skipping security protections (e.g., anti-virus software), although that work did not include any experimental validation. Schroeder [53] conducted a lab-based survey of IT officers in the U.S. military and found that prospect theory predicted hypothetical decisions about investment in information security. Sawicka and Gonzalez [51] explored the extent to which prospect theory can explain behavioral dynamics in IT-based work environments; they found the model matched choices observed in a short experimental run, but that it was not likely to account accurately for behavior over longer time periods. Sanjab et al. [50] explored how the decision weight function and value function impact principals' decisions in adversarial games in the context of attacks on Unmanned Aerial Vehicles (UAVs); they found that these subjective functions led to the adoption of riskier strategies, which cause delays in delivery. Most recently, Qu et al. [48] investigated the reference-dependence effect and the pseudocertainty effect in the context of two-factor authentication; they found that both effects explained whether or not users choose to enable two-factor authentication for a game in a laboratory setting. However, other security decisions—notably including password selection—have not been previously studied.

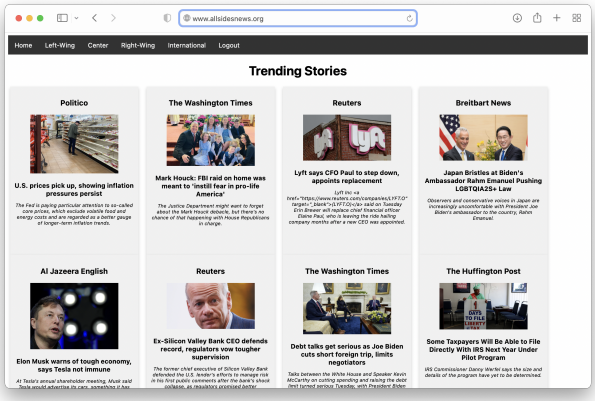
Applications of Prospect Theory to Privacy. In 2007, Acquisti et al. posited that several prospect theory effects—notably ambiguity aversion—might significantly impact privacy decision making [2]. Follow-up work found that people were more willing to sell personal information than to buy back previously-disclosed information [3, 27], and that the framing of notices affected whether or not users disclosed personal information in a survey [4]. Chloe et al. [14] also found that visual signals of an app's trustworthiness were affected by framing, but found that *positively* framed signals were more effective at nudging users away from low-privacy apps. More recent work has looked at developing and validating a theory for how context and personality affect decisions about disclosing personal information [5] and at the mechanism-design problem of how to calibrate noise in privacy-preserving mechanisms [40, 41].



(a) Account creation page



(b) Example interactive prompt



(c) Website home page

Figure 2: Screenshots of the account creation process on the example site

4 Methodology

To investigate how well prospect theory effects apply as a descriptive model of password selection, we conducted a pair of online user studies to evaluate the impact of two prospect theory effects—the source-dependence effect and the reference-dependence effect—on password selection decisions. These studies also explored the impact of feedback and mental models on password selection.

4.1 Experimental Setup

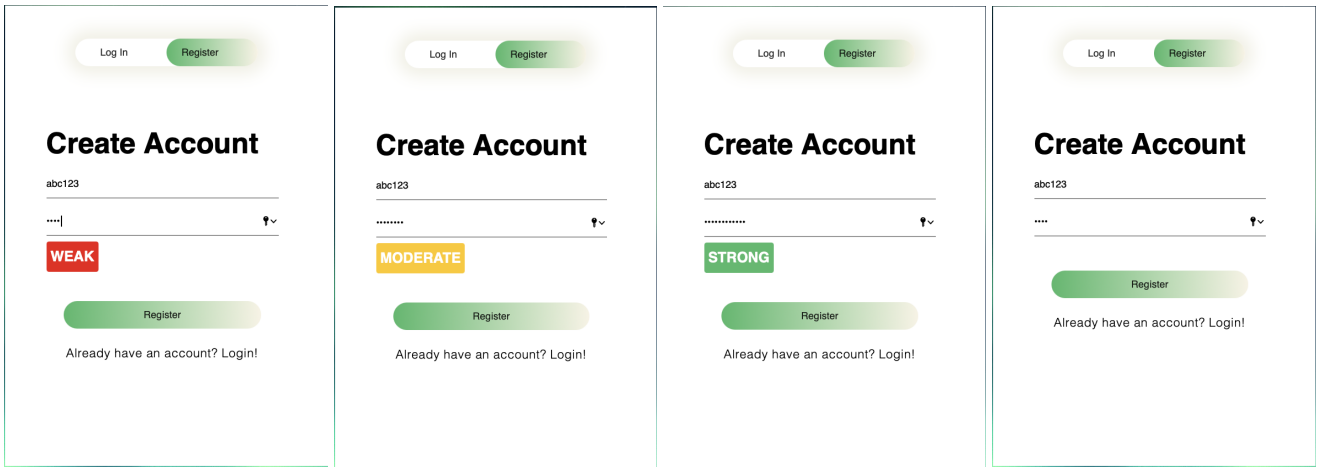
We developed an experimental aggregated news site that is accessible only to authenticated users. When visiting for the first time, each user is pseudorandomly assigned to a condition based on a hash of their current IP address. The user is then redirected to a condition-specific version of the account creation page (Figure 2a).

The initial account creation page had two different versions:

1. *Password Meter:* In these conditions, the password strength is classified in real time using the `zxcvbn` password strength estimator [70], and this information is displayed to the user by a password meter. Each password is classified as weak if it has a `zxcvbn` total score of 0 or 1 and moderate if the password has a total score of 2. Passwords with a total score of 3 or 4 are considered strong. Screenshots showing examples of how this meter looks with weak, moderate, and strong passwords are shown in Figures 3a, 3b, and 3c.

2. *No Meter:* In this condition, no information is displayed to the user about the strength of their password. This condition is shown in Figure 3d.

All participants in User Study 1 and most participants in User Study 2 saw an account creation page with a password meter. To provide a baseline for exploring the impact of feedback on password selection decisions, User Study 2 also included a condition with no meter.



(a) Meter (weak password) (b) Meter (moderate password) (c) Meter (strong password) (d) No Meter

Figure 3: Screenshots depicting the initial account creation page in different conditions with different strength passwords.

After initially selecting a password, users who select a strong password are redirected to the home page of the aggregated news site (Figure 2c). Users who select a weak or moderate password are instead presented with an interactive prompt that states that weak (resp., moderate) passwords put their account at risk and asks whether they would like to choose a stronger password (Figure 2b).

This prompt was presented using one of four possible wordings:

1. *Vague Prompt:* The password you selected is ⟨*strength*⟩. Would you like to choose a stronger password?

2. *Specific Prompt:* ⟨*strength*⟩ passwords can be guessed or learned by attackers in ⟨*time*⟩, which may lead to the loss of personal information, including credit card info, and identity theft. Would you like to choose a stronger password?

3. *Moderate Prompt:* The password you selected is ⟨*strength*⟩. This password can be guessed by attackers in ⟨*time*⟩. Would you like to choose a stronger password?

4. *Moderate Prompt + Suggestions:* The password you selected is ⟨*strength*⟩. This password can be guessed by attackers in ⟨*time*⟩. Would you like to choose a stronger password? Suggestions to improve password: ⟨*suggestions*⟩

Here ⟨*strength*⟩ is the classification based on the `zxcvbn` total score of the password the user submitted: "Weak" or "Moderate". (Recall that users who submit a strong password are authenticated immediately and are not presented with a prompt.) For the moderate and specific prompts, ⟨*time*⟩ is the `zxcvbn` estimate for how long it would take to crack the password with an offline guessing attack if passwords are hashed and salted using a slow hashing algorithm with a moderate work factor (e.g., bcrypt, scrypt, or PBKDF2). We used the human-readable text generated by `zxcvbn`, for example, "less than a second", "20 seconds", or "5 minutes". ⟨*suggestions*⟩ used the the natural-language text suggestions generated by `zxcvbn`, which provided password-specific suggestions such as "Avoid repeated words and characters", "Add another word or two. Uncommon words are better", and "Predictable substitutions like '@' instead of 'a' don't help very much". Participants in User Study 1 were shown either a vague prompt or a specific prompt. All participants in User Study 2 were shown a moderate prompt, with some conditions including suggestions and some not.

As shown in Figure 2b, the prompt has two buttons: one to go back (and choose a different password) and one to continue creating the account with the current password. This pair of buttons is labeled with one of three possible framings:

1. *Positive Framing:*
   Go Back: Choose a stronger password to reduce the risks of financial loss and identity theft
   Continue: Create account with current password

2. *Neutral Framing:*
   Go Back: Yes
   Continue: No

3. *Negative Framing:*
   Go Back: Choose a stronger password
   Continue: Ignore potential risks of financial loss and identity theft and create account with current password

The identified threats—financial loss and identity theft—were selected to maximize the appearance of risk within the password selection decision. However, we believe these risks are appropriate to this context. Prior work has established that password reuse attacks—in which attackers use leaked cre-

Study	Meter?	Prompt	Framing
1	Yes	Vague	Positive
1	Yes	Vague	Neutral
1	Yes	Vague	Negative
1	Yes	Specific	Positive
1	Yes	Specific	Neutral
1	Yes	Specific	Negative
2	No	Moderate	Neutral
2	Yes	Moderate	Positive
2	Yes	Moderate	Neutral
2	Yes	Moderate	Negative
2	Yes	Moderate+Suggestions	Positive
2	Yes	Moderate+Suggestions	Neutral
2	Yes	Moderate+Suggestions	Negative

Table 1: Conditions included in the two user studies.

dentials from low-value accounts such as news websites to attempt to access high-value accounts such as back accounts and emails—commonly occur online [9]. These attacks, which take advantage of the common practice of reusing credentials across multiple websites—are suspected behind several high-profile account compromises [31].

Each condition is defined by its meter setting (no meter vs. password meter), the wording of its interactive prompt (vague, specific, moderate, or moderate + suggestions), and its framing (postive, neutral, or negative). The conditions included in each of our two user studies are summarized in Table 1.

Users who elect to continue are redirected to the site homepage. Users who choose to go back stay on the account creation page until they select and submit a second password; they are then redirected to the site home page (no matter how strong their second selected password is). After spending a short time on the site, study participants returned to Qualtrics and completed a follow-up survey. Participants in the second user study were also asked to re-authenticate on the website after completing the survey. The full sets of questions for the follow-up surveys are provided in Appendix A and Appendix B.

The precise information recorded varied between the two studies. In the first user study, the experimental site logged the coarse-grained strength of each user's initial password choice (weak, moderate, or strong), how they interacted with the interactive prompt (if applicable), and the coarse-grained strength of their second password choice (if applicable). In the second user study, the site additionally logged the number of guesses it would take to crack each password (as estimated by `zxcvbn`), the length of each password, and (for users who selected a second password) the edit distance between the two passwords. In the second study, the site also stored the salted hash of the final password selected; this information was deleted after data collection was complete. Due to ethical and security concerns, we did not record any plaintext passwords.

4.2 Participant Recruitment

We recruited participants for both user studies online. All participants were presented with a consent form that informed them about what data would be collected and how that data would be used; only people who consented to these practices participated in a study. Our user studies, including all consent forms and survey instruments, were reviewed and approved in advance by the Pomona College Institutional Review Board.

User Study 1. For our first user study, participants were recruited through Amazon Mechanical Turk. Participation was restricted to United States residents who had completed at least 50 HITs with an approval rate of at least 95%.

The task was advertised as beta-testing an aggregated news site. Each participant was asked to (1) spend 1-2 minutes exploring the website as they would normally behave as an Internet user, (2) enter the unique confirmation code displayed when they visited the site, and (3) complete the follow-up survey questions. To avoid the appearance of collecting any personal information, users were given an email address to use during account creation.

Participants who did not enter a valid confirmation code, for whom we had no recorded log data, or who submitted irrelevant or incoherent responses to our free-response attention check question were excluded from the study. The 762 participants who completed the full study were compensated $1.20. Median completion time for this study was 5.15 minutes.

User Study 2. Due to increasing concerns about the external validity of studies conducted on Amazon Mechanical Turk [57], we elected to recruit participants for our second user study through Prolific. We recruited a gender-balanced, U.S. sample; following the methodology of Tang et al. [57], we did not further restrict participation.

Because Prolific is exclusively a platform for conducting studies, we advertised our second user study as a study about how people interact with websites instead of framing it as beta testing. Participants were given the same instructions as in User Study 1. However, since Prolific provides all users with an anonymous email tied to their Prolific ID—through which messages can be sent to the internal Prolific messaging system—we asked participants to use that email address to create their account.

We applied the same exclusion criteria in both studies. The 607 participants who successfully completed User Study 2 were compensated $2.50. The median completion time for the full task was 7.07 minutes. The higher compensation compared to User Study 1 was due to a longer estimated completion time combined an increase in California's minimum wage between the two studies.

The demographics of our study populations are summarized in Table 2.

Demographic		Study 1	Study 2	U.S.
Age	18-24	7.3	19.1	11.9
	25-34	33.6	35.3	17.9
	35-44	33.4	23.4	16.4
	45-59	19.8	16.5	24.4
	60-74	5.7	5.4	20.6
	75+	0.2	0.2	8.8
Race	White	76.1	77.8	74.4
	Black	10.4	11.4	13.9
	Asian	12.1	12.2	6.6
	Native Am.	3.1	2.4	1.5
	Other	2.3	2.3	5.2
Gender	Male	53.4	49.9	48.7
	Female	45.6	46.8	51.3
	N.B./other	1.0	3.3	-

Table 2: Study population demographics compared to the demographics of the United States, as published in the American Community Survey (ACS).

4.3 Study Limitations

In this study, our participants selected passwords for an experimental news site—one that they would likely not access or use beyond the scope of the study—rather than select passwords for a real-world account. This lack of realism might have affected password selection decisions. While prior work has shown that people select similar passwords in online studies compared to passwords selected for real accounts [21, 43], that work was conducted 10 years ago and focused exclusively on Amazon Mechanical Turk.

Moreover, this work looks specifically at the impact of rephrasing and re-framing risks incurred by password selection decisions. However, participants did not use personal email addresses in our studies; participants might therefore not have felt that their password selection decision put them at risk.

Finally, the particular threats emphasized in the experimental design—threats of financial loss and identity theft—might not have resonated with all participants, since the experimental website was an aggregated news site that did not direct collect and personal or financial information.

We discuss the validity of our results further in Section 6.

4.4 Hypotheses

To explore how prospect-driven interventions impact password selection, we identified and evaluated six hypotheses.

Source-dependence effect. When presented with the vague prompt, a user is required to evaluate options in the presence of multiple different sources of risk: in addition to reasoning about how likely it is that an attacker would target this site or this user, the user must evaluate uncertainties about how hard it would be for an attacker to guess their password and about what the potential consequence of password compromise might be. When presented with the specific prompt, some of these uncertainties—in particular how hard it would be for an attacker to guess their password and what attackers might do after they have learned a user's password—are eliminated in favor of more concrete risks.

The source-dependence effect observes that users evaluate different types of risk differently, and in particular that ambiguities are evaluated differently than more concrete risks. We therefore hypothesize that users will evaluate the the option to continue with their current (weak or moderate) password more negatively when presented with the specific prompt than with the vague prompt, resulting in stronger password selection after interacting with the specific prompt compared to the vague prompt.

Hypothesis 1: *Users' password selection decisions exhibit the* source-dependence effect*, that is users assigned to the specific prompt conditions are more likely to strengthen their password and will ultimately select stronger passwords compared to users assigned to the vague prompt conditions.*

Reference-dependence effect. In the conditions with positive framing, the option to go back is labeled as "Choose a stronger password to reduce the risks of financial loss and identity theft". By emphasizing the benefits of going back, this framing implicitly nudges the user to consider the option to continue as the neutral reference point and the option to go back as a choice with higher utility relative to that reference point. By contrast, the negative framing emphasizes the loss of utility ("potential risks of financial loss and identity theft") associated with continuing with the current password, thereby implicitly nudging the user to treat the option to go back as the neutral reference point and to evaluate continuing as a loss of utility relative to that reference point.

The reference-dependence effect implies that this difference in framing will cause users assigned to a positive framing condition to evaluate the difference between going back (i.e., choosing a stronger password) and continuing (i.e., submitting a weak password) as a positive *gain* in utility, whereas users assigned to a negative framing condition will evaluate the difference between continuing (i.e., submitting a weak password) and going back (i.e., choosing a stronger password) as a *loss* of utility. The loss aversion effect suggests that the subjective value function is steeper for (relative) losses than for (relative) gains. We therefore hypothesize that users will evaluate the option to continue with the current (weak) password more negatively in the negative framing conditions than the positive framing conditions—even though the two options have the same absolute utility in all conditions—resulting in stronger passwords selected after interacting with the negative framing prompt compared to the neutral and positive framing prompts.

Hypothesis 2: *Users' password selection decisions exhibit the* reference-dependence effect*, that is users assigned to a negative framing condition—which frames going back as the neutral baseline and continuing as a loss relative to that baseline—are more likely to strengthen their password and will ultimately select stronger passwords compared to users assigned to neutral or positive framing conditions.*

Feedback and Suggestions. Even after users decide to improve the strength of their password, the ability to successfully do so depends on knowing what constitutes a stronger password. The presence of a real-time, interactive password meter—which gives users course-grained feedback about the strength of their password—is one way to enable users to discover what might constitute a stronger password. Another approach would be to provide users with concrete suggestions for how they might strengthen their password.

Hypothesis 3: *Users who are shown a real-time password meter—which rates the current strength of their password—are more likely to strengthen their password and will ultimately select a stronger password compared to users who have no real-time information about their passwords strength.*

Hypothesis 4: *Users who are given concrete suggestions for how to improve their password are more likely to strengthen their password and will ultimately select a stronger password compared to users who are not shown suggestions.*

Mental Models of Hacking. Our user study concluded with a series of questions about participants' mental models of hacking and password security. One question we asked was who participants believe are the primary targets of password stealing attacks. Drawing on Wash's taxonomy of hacker mental models [68], we provided three possible answer: hackers target everyone equally, hackers primarily target rich people, and hackers primarily target users with special privileges (e.g., system administrators). We hypothesized that users who believe that hackers target everyone equally will consider themselves to be a more likely target compared to users with other mental models and will therefore be more sensitive to risks associated with password compromise.

Hypothesis 5: *Users who believe everyone is equally likely to be targeted by a password stealing attack will be more likely to strengthen their password and will ultimately select stronger passwords.*

We also asked participants questions designed to understand how they evaluated password-related risks. In particular, we asked how likely they believed a password attack would be to compromise their password if they selected a weak (resp., moderate, strong) password. We hypothesized that users' beliefs about risks associated with passwords would correlate with users' password selection decisions.

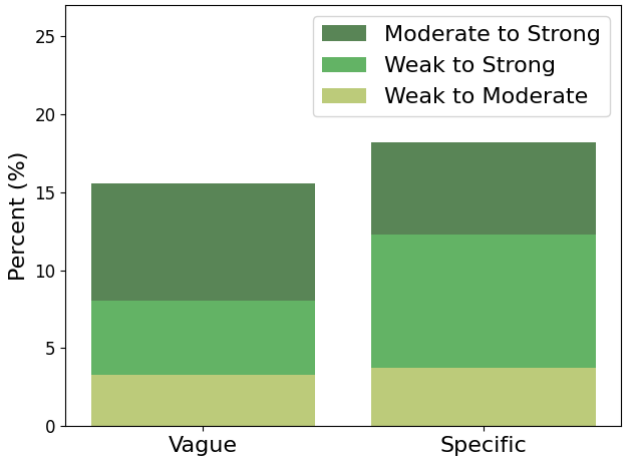


Figure 4: Percentage of users who improved password strength after interacting with vague and specific prompts.

Hypothesis 6: *Users who believe that weak passwords are more likely to be guessed by attackers will be more likely to initially choose a strong password, will be more likely to strengthen their password after seeing an interactive prompt, and will be more likely to ultimately choose a strong password.*

5 Results

To evaluate our hypotheses, we focused on users who initially selected weak or moderate passwords (i.e., users who saw the interactive prompt), and measured how many of those users (1) decided to go back and (2) selected a stronger password. We used χ^2-contingency tests to test for statistically significant differences.

5.1 Source-dependence Effect

To evaluate Hypothesis 1, we compared behavior in the conditions with vague wording to conditions with specific wording using data collected in User Study 1. We did not include conditions with moderate wording to avoid introducing confounding effects due to differences between study populations. We found that the specificity of the prompt had no significant effect on password selection. Of the users who saw a vague prompt, 15.6% opted to go back and ultimately selected a stronger password, compared to 18.2% of users who saw the specific prompt. This difference, depicted in Figure 4, was not statistically significant ($\chi^2 = .3$, $p = .573$).

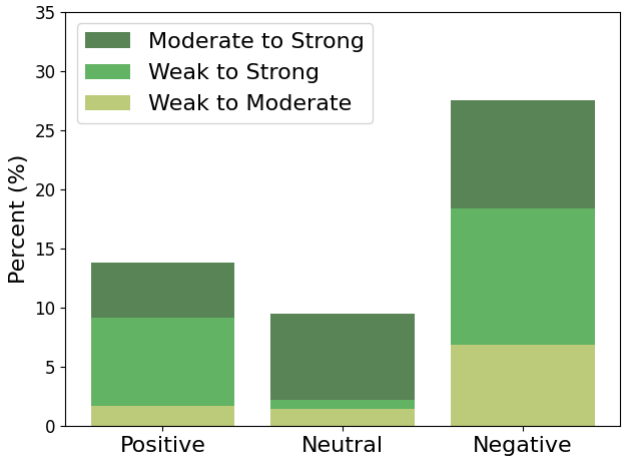


Figure 5: Percentage of users who improved password strength after seeing prompts with various framings.

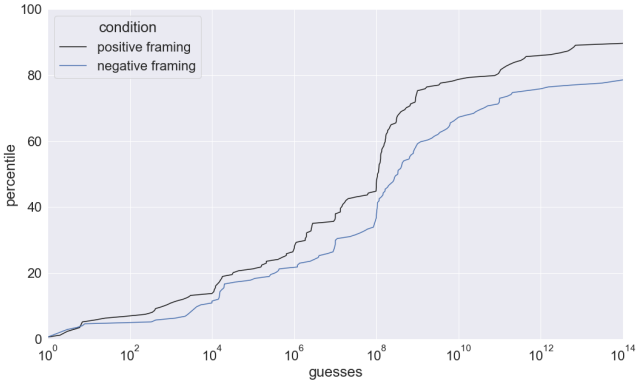


Figure 6: Strength of final password chosen by users who saw prompts with positive and negative framings, as measured by the percentile of final passwords in each condition that could be cracked with various numbers of guesses. Fewer passwords cracked corresponds to stronger passwords.

This negative result might be an indication that the source-dependence effect does not apply in the context of password selection decisions. However, it is also possible that the language of our prompts was insufficient to transfer uncertainty-based risk into probability-based risk in a manner that would trigger the source-dependence effect. Finally, it is possible that many of our users simply did not read the prompt, precluding the possibility of observing statistically significant effects due to the source-dependence effect.

Regardless of the underlying mechanism, these results suggest that utilizing more specific language about the nature of risks due to weak passwords—including notifying users of how long it would take an attacker to crack a password—is not an effective way to nudge users to select stronger passwords.

5.2 Reference-dependence Effect

To evaluate Hypothesis 2, we measured how many people strengthened their password after an intervention with negative framing compared to an intervention with positive framing (resp. neutral framing) using data collected in User Study 1. We conducted pairwise χ^2 tests to determine whether differences were significant. 25.9% of participants who saw a banner with negative framing went back and improved the strength of their password. This was significantly higher than the 14.2% who improved their password after seeing a banner with positive framing ($\chi^2 = 4.9$, $p = .027$) and the 9.5% who improved their password after seeing a banner with neutral framing ($\chi^2 = 11.5$, $p < .001$). There was no significant difference between the neutral framing and positive framing conditions ($\chi^2 = 1.0$, $p = .323$).

To validate that an interaction with negative framing improves password strength, we compared fine-grained strength—as measured by the estimated number of guesses it would take to crack that password—of the final password for our positive and negative framing conditions in User Study 2. We found that participants who saw a prompt with negative framing ultimately selected stronger passwords—i.e., passwords that would take more guesses for an attacker to crack—relative to participants who saw a prompt with positive framing (Figure 6).

These results suggest that the reference-dependence effect occurs in the context of password selection decisions. While further work will be required to validate this result in real-world systems, prior work has found that the results of password studies conducted online generally do extend to real-world systems [21]. The insight that the reference-dependence effect models users' password selection decisions therefore provides guidance for how authentication mechanisms designers might prescriptively enhance security: by adding a confirmation page and framing the option to go back and select a strong password as the "baseline" (and framing the option to continue with a weak or moderate password as a loss of utility relative to that baseline), we might be able to nudge users to enhance the security of their accounts by selecting significantly stronger passwords.

5.3 Feedback and Suggestions

To evaluate Hypothesis 3 and 4, we analyzed data from User Study 2 and compared conditions with a password meter and no suggestions to (1) our condition with no password meter and (2) our conditions that provided concrete suggestions for how to strengthen the initial password. In all cases, the fraction of participants who successfully strengthened their password was 22-26% (shown in Figure 7). There was no significant difference from removing the password meter ($\chi^2 < .1$, $p = .994$) or adding suggestions ($\chi^2 = .3$, $p = .606$).

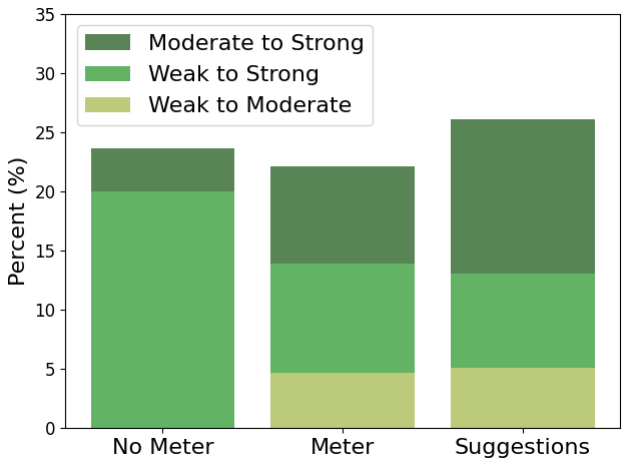


Figure 7: Percentage of users who improved passwords based on information provided about how to improve it.

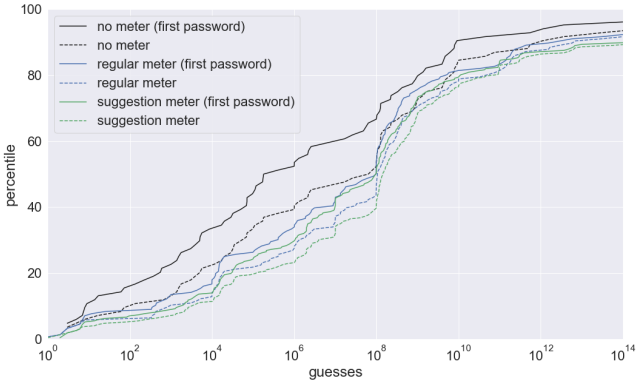


Figure 8: Impact of feedback and suggestions on password strength, measured by the percentile of passwords that could be cracked with various numbers of guesses. Fewer passwords cracked corresponds to stronger passwords.

To further understand this surprising negative result, we looked at data collected about fine-grained password strength—as defined by the number of guesses estimated to crack the password—both for the initial password selected and for the final password selected by participants in User Study 2. Like much prior work, we found that providing a password meter improved the strength of the initial password selected. Although final passwords selected in the no-meter condition were ultimately weaker than those selected in the conditions with a meter or a meter and suggestions, this distinction seems to be due to those weaker initial passwords rather than decreased ability to improve passwords without a password meter. We did not observe any significant differences between the conditions with and without suggestions, either in the strength of the initial password or in the strength of the ultimate password selected. These results are depicted in Figure 8.

Although we did not record plaintext passwords in either study, we did record password length and the edit distance between the two passwords (for people who selected a second password) for participants in User Study 2. We looked at this data to better understand the types of changes people made to their passwords. Overall, we found that 71% of people who decided to go back were successful at improving password strength. 8% stuck with their original password despite going back. 5% made small edits (defined as and edit distance of 3 or less) that did not improve strength. 13% made large edits that did not yield a stronger password; many of these large edits constituted selecting a completely new password. These results suggest that suggestions and feedback might be helpful for a minority of users; however, most people appear to already know how to strengthen their password and simply have to make the decision to do so. Definitively determining whether suggestions make prospect-driven interventions more effective will therefore require further work with larger sample sizes.

5.4 Mental Models

In our follow-up survey, we asked about users' mental models of password risks in order to explore whether there was a correlation between how users thought about password attacks and how users responded to our interactive prompts.

Hacking Targets. We asked participants in both user studies to identify who they thought hackers would target: everyone equally, primarily rich people, or primarily privileged users (e.g., system administrators). Overall, we found that 69.9% of participants believed that hackers target everyone equally and anyone is equally likely to have their password stolen, 10.6% of participants believed that hackers primarily target rich people, and 15.7% of participants believed that hackers primarily target privileged accounts (Figure 9).

A small number of participants opted instead to provide a free-form response. Some of these responses identified alternate groups as primary targets, including "gullible people", "weak links", and "older people". Other responses provided more nuanced variants of the options provided, e.g., "It depends on the hacker. Botnets attack everyone while social engineering attacks focus on special privileges" or identified all of the above as the best description of who is likely to be the target of a password attack.

Users who believed that hackers primarily target administrators during such attacks were significantly less likely to improve the strength of their password after exposure to the interactive prompt compared to users who believed that everyone is targeted equally ($\chi^2 = 4.9$, $p = .027$). We believe

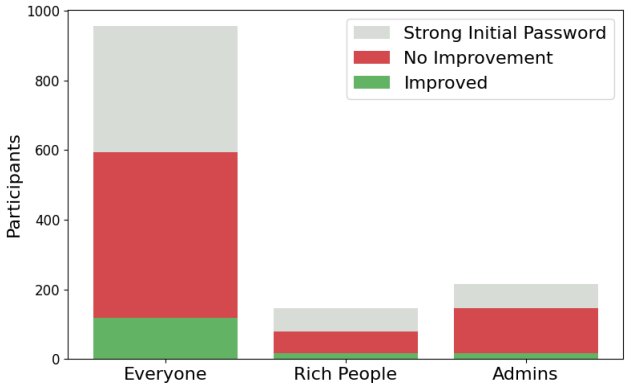


Figure 9: Password selection decisions broken down by perceived target of attacks.

this difference occurs because users with this mental model are less likely to believe they will be the target of an attack. To our surprise, users who believed that hackers target everyone equally were no more likely to improve the strength of their password compared to users who believed that hackers primarily target rich people ($\chi^2 = .1$, $p = .797$). This might be due to the fact that Americans consistently underestimate income inequality [17, 46, 47] and the income of top earners relative to the median worker [34] and thus might consider themselves to be a high-priority target even if they hold that mental model. Prior work has also found that participants recruited through Mechanical Turk and Prolific are more highly educated than the overall population [49, 57], a demographic that correlates with income and wealth.

Risk Evaluation. We also asked survey participants to rate how likely an attack would be to successfully compromise a password if a user selected (1) a weak password, (2) a moderate password, or (3) a strong password. We found that 88.1% of participants considered a weak password to be somewhat or very likely to be successfully attacked, compared to 53.7% of participants for a moderate password and 17.8% of participants for a strong password. These responses, which are depicted in Figure 10, were statistically significantly different between all the different password strengths ($\chi^2 \geq 382.4$, $p < .001$). These results suggest that most users believe that stronger passwords are in fact less likely to be vulnerable to password guessing attacks. However, there was no significant correlation between whether a user believed stronger passwords had less risk of being compromised and whether that user improved the strength of their password after seeing the interactive prompt.

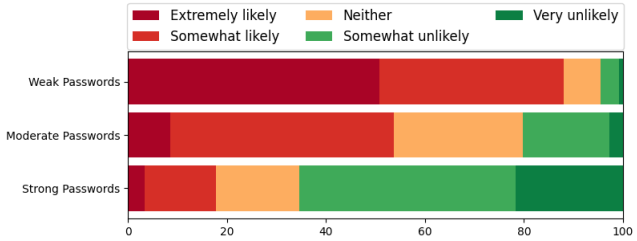


Figure 10: Perceptions of how likely a password guessing attack is to succeed based on the strength of a user's password.

6 Discussion and Limitations

Our results suggest that it might be possible to significantly improve the strength of user-chosen passwords by leveraging insights from prospect theory—in particular the reference-dependence effect—through a negatively-framed interactive prompt after users select an initial password. However, further work and careful consideration will be required to determine whether and how we should leverage these effects.

Ecological Validity. The major limitation of this work arises from the fact that we recruited participants through online crowdsourcing platforms to select passwords for an experimental account. Prior work has found that online study participants select slightly weaker passwords in experimental settings compared to real accounts. For example, one study found that 44.0% of users selected guessable passwords for their real account compared to 47.5% of Mechanical Turk users who were asked to select a password for an experimental study account given identical constraints [43]; in our first user study (also conducted on Mechanical Turk), we similarly found that 47.0% of our users initially selected a weak password (Figure 11). Despite these slight discrepancies, prior work has found that results from laboratory and online studies about passwords correspond to patterns in behavior for real accounts [21].

In addition to general threats to validity common to all online password studies, our focus on prospect theory—and the resulting need for participants to feel that their decisions might incur risk—introduces additional threats to validity. For ethical reasons, we did not collect any personal information (including personal email addresses) and instead had participants use dummy email addresses (User Study 1) or anonymous Prolific addresses (User Study 2), so participants might have realized that there was no actual risk incurred. This (lack of) realism might have influenced participants' decisions. Moreover, the choice of language in the warnings might have influenced how people reacted; some people might have disbelieved that weak passwords on news websites might incur financial risk. To explore how such confounding effects

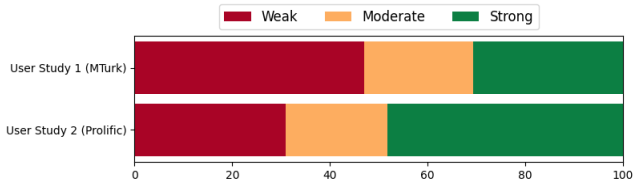


Figure 11: Strength of passwords initially selected by users.

might have affected our results, we asked participants in User Study 2 why they made the decision they made. About half of our participants responded in ways that suggested they were acting as though there were real risks (e.g., "Because I don't want to be at risk", "I didn't want my password to be too easy to guess", and "Its better to be safe than sorry"). However, other participants mentioned lack of realism (e.g., "This was not a real account and will not be using it again") or disbelief about the alleged risks (e.g., "Because I don't have any banking information on the website").

These results emphasize the importance of validating these effects (and their magnitude) in real-world contexts with actual risks. However, we hypothesize that the observed reference-dependent effect will extend to real-world password selection decisions, perhaps even with a larger effect size.

MTurk vs. Prolific. Recent work found that external validity of security and privacy surveys on Mechanical Turk has degraded over the last five years and that surveys conducted through Prolific now have higher external validity [57]. However, to our knowledge, this is the first work to conduct comparable experimental studies on both platforms.

Most results from our studies were consistent. People behaved similarly in both studies, e.g., significantly more people improved their password after seeing an intervention with negative framing (14.2% in User Study 1, 12.8% in User Study 2) than after seeing an intervention with negative framing (25.9% in User Study 1, 33.3% in User Study 2). Survey responses were also similar; e.g., 70.7% of participants in User Study 1 believed that attackers target everyone equally compared to 68.7% in User Study 2.

However, there was one notable difference between the studies: the strength of password people initially selected. Only 30.9% of participants in User Study 2 selected a weak password when presented with the same account creation interface used in User Study 1 compared to 47.0% in User Study 1 (Figure 11). This difference, which is statistically significant ($\chi^2 = 40.5$, $p < .001$) might be due to population differences between the two online platforms. Alternatively, it might be due to differences in how the study was presented. Since Prolific is a dedicated research study platform, participants in User Study 2 were aware all along that they were participating in a study rather than beta testing a website, which might have resulted in users selecting less realistic passwords. On the other hand, participants recruited through Prolific used a valid (albeit anonymous) email to create their account, perhaps resulting in participants selecting more realistic passwords. Differences between the observed rate of weak passwords in User Study 2 and that observed by prior work with real-world accounts might be symptomatic of low validity or might reflect temporal shifts in password selection behavior. Further research will be required to quantify the validity of experimental studies conducted on Prolific compared to Mechanical Turk and to validate password selection behavior in such studies today.

Memorability. The risk of account compromise due to password cracking and other attacks is not the only risk that users consider when selecting a password: users also need to weigh risks associated with other factors such as memorability. Forgetting a password is inconvenient in the best case; in the worst case, users can lose access to accounts. Future work will be required to determine the effect of framing on the memorability of passwords that users select.

Concerns about memorability might motivate users to employ memory-assistance techniques. This could lead to improved security practices—such as increased adoption of password managers—or to bad security practices—such as writing down passwords and leaving them in accessible locations. Further work will be required to evaluate the impact of framing on these other password-related practices.

Ethical Considerations. Leveraging the reference-dependence effect through negative framing of decisions has the potential to enhance security by encouraging users to adopt stronger passwords. However, this effect is an example of nudging [1]. While nudging is often associated with manipulative design elements that nudge users to make decisions that are inimical to their interests [10, 11, 15, 26], nudging can also be used towards making decisions that the mechanisms designer views as "better", a form of nudging sometimes called *soft paternalism* [22, 35, 52, 58]. Since nudging inherently leverages subconscious patterns in human behavior, care and consideration will be required to ensure that any prescriptive application of nudging and prospect theory effects with real-world impact—including leveraging the referenced-dependence effect to improve password selection—is handled ethically and responsibly.

Recommendations. Based on our results, we recommend adoption of password selection interfaces that present strong passwords as the default choice and follow-up prompts that emphasize risks associated with weak passwords. However, care will be required to ensure that this is ethically and effectively done without compromising memorability or other priorities. We recommend further research to to validate these results within real-world deployments and to ensure that the potential benefits to security outweigh any potential harms.

References

[1] Alessandro Acquisti, Idris Adjerid, Rebecca Balebako, Laura Brandimarte, Lorrie Faith Cranor, Saranga Komanduri, Pedro Giovanni Leon, Norman Sadeh, Florian Schaub, Manya Sleeper, Yang Wang, and Shomir Wilson. Nudges for privacy and security: Understanding and assisting users' choices online. *ACM Computing Surveys*, 50(3):1–41, 2017.

[2] Alessandro Acquisti, Stefanos Gritzalis, Costos Lambrinoudakis, and Sabrina di Vimercati. *What can behavioral economics teach us about privacy?* Auerbach Publications, 2007.

[3] Alessandro Acquisti, Leslie K. John, and George Loewenstein. What is privacy worth? *The Journal of Legal Studies*, 42(2):249–274, 2013.

[4] Idris Adjerid, Alessandro Acquisti, Laura Brandimarte, and George Loewenstein. Sleights of privacy: Framing, disclosures, and the limits of transparency. In *Symposium on Usable Privacy and Security*, pages 1–11, 2013.

[5] Gaurav Bansal, Fatemeh Mariam Zahedi, and David Gefen. Do context and personality matter? Trust and privacy concerns in disclosing private information online. *Information & Management*, 53(1):1–21, 2016.

[6] Nicholas Barberis and Ming Huang. Stocks as lotteries: The implications of probability weighting for security prices. *American Economic Review*, 98(5):2066–2100, 2008.

[7] Nicholas C. Barberis. Thirty years of prospect theory in economics: A review and assessment. *Journal of Economic Perspectives*, 27(1):173–96, 2013.

[8] Levon Barseghyan, Francesca Molinari, Ted O'Donoghue, and Joshua C Teitelbaum. The nature of risk preferences: Evidence from insurance choices. *American Economic Review*, 103(6):2499–2529, 2013.

[9] Bitglass. Where's your data? https://pages.bitglass.com/Bitglass_Where_is_your_Data_Report.html, 2016.

[10] Christoph Bösch, Benjamin Erb, Frank Kargl, Henning Kopp, and Stefan Pfattheicher. Tales from the dark side: Privacy dark strategies and privacy dark patterns. *Proc. Priv. Enhancing Technol.*, 2016(4):237–254, 2016.

[11] Harry Brignull and Alexander Darlo. Dark patterns. *Dark Patterns*, 2019.

[12] Colin Camerer, Linda Babcock, George Loewenstein, and Richard Thaler. Labor supply of New York City cabdrivers: One day at a time. *The Quarterly Journal of Economics*, 112(2):407–441, 1997.

[13] Xavier De Carné De Carnavalet and Mohammad Mannan. A large-scale evaluation of high-impact password strength meters. *ACM Transactions on Information and System Security (TISSEC)*, 18(1):1–32, 2015.

[14] Eun Kyoung Choe, Jaeyeon Jung, Bongshin Lee, and Kristie Fisher. Nudging people away from privacy-invasive mobile apps through visual framing. In *IFIP Conference on Human-Computer Interaction*, pages 74–91. Springer, 2013.

[15] Gregory Conti and Edward Sobiesk. Malicious interface design: Exploiting the user. In *Proceedings of the 19th International Conference on World Wide Web*, pages 271–280, 2010.

[16] Vincent P. Crawford and Juanjuan Meng. New York City cab drivers' labor supply revisited: Reference-dependent preferences with rational-expectations targets for hours and income. *American Economic Review*, 101(5):1912–32, 2011.

[17] Shai Davidai. Why do Americans believe in economic mobility? Economic inequality, external attributions of wealth and poverty, and the belief in economic mobility. *Journal of Experimental Social Psychology*, 79:138–148, 2018.

[18] Anzo DeGiulio, Hanoom Lee, and Eleanor Birrell. "Ask app not to track": The effect of opt-in tracking authorization on mobile privacy. In *Emerging Technologies for Authorization and Authentication (ETAA)*, pages 152–167. Springer, 2021.

[19] Stephen G. Dimmock and Roy Kouwenberg. Loss-aversion and household portfolio choice. *Journal of Empirical Finance*, 17(3):441–459, 2010.

[20] Serge Egelman, Andreas Sotirakopoulos, Ildar Muslukhov, Konstantin Beznosov, and Cormac Herley. Does my password go up to eleven? The impact of password meters on password selection. In *SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 2379–2388, 2013.

[21] Sascha Fahl, Marian Harbach, Yasemin Acar, and Matthew Smith. On the ecological validity of a password study. In *Proceedings of the Ninth Symposium on Usable Privacy and Security*, pages 1–13, 2013.

[22] Bijan Fateh-Moghadam and Thomas Gutmann. Governing [through] autonomy. The moral and legal limits of "soft paternalism". *Ethical Theory and Moral Practice*, 17(3):383–397, 2014.

[23] Milton Friedman and Leonard J. Savage. The utility analysis of choices involving risk. *Journal of political Economy*, 56(4):279–304, 1948.

[24] Roland G. Fryer Jr, Steven D. Levitt, John List, and Sally Sadoff. Enhancing the efficacy of teacher incentives through loss aversion: A field experiment. Technical report, National Bureau of Economic Research, 2012.

[25] Maximilian Golla and Markus Dürmuth. On the accuracy of password strength meters. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, pages 1567–1582, 2018.

[26] Colin M. Gray, Yubo Kou, Bryan Battles, Joseph Hoggatt, and Austin L. Toombs. The dark (patterns) side of UX design. In *Proceedings of the 2018 SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 1–14, 2018.

[27] Jens Grossklags and Alessandro Acquisti. When 25 cents is too much: An experiment on willingness-to-sell and willingness-to-protect personal information. In *Workshop on Economics of Information Security*, 2007.

[28] Paul Heidhues and Botond Kőszegi. Regular prices and sales. *Theoretical Economics*, 9(1):217–251, 2014.

[29] Tanjim Hossain and John A. List. The behavioralist visits the factory: Increasing productivity using simple framing manipulations. *Management Science*, 58(12):2151–2167, 2012.

[30] Wei-Yin Hu and Jason S. Scott. Behavioral obstacles in the annuity market. *Financial Analysts Journal*, 63(6):71–82, 2007.

[31] David Jaeger, Chris Pelchen, Hendrick Graupner, Feng Cheng, and Christoph Meinel. Analysis of publicly leaked credentials and the long story of password (re-) use. *Hasso Plattner Institute, Universidad de Potsdam. Disponible en https://bit. ly/2E7ZT01*, 2016.

[32] Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–292, 1979.

[33] Patrick Gage Kelley, Saranga Komanduri, Michelle L Mazurek, Richard Shay, Timothy Vidas, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, and Julio Lopez. Guess again (and again and again): Measuring password strength by simulating password-cracking algorithms. In *IEEE Symposium on Security and Privacy*, pages 523–537, 2012.

[34] Sorapop Kiatpongsan and Michael I. Norton. How much (more) should CEOs make? A universal desire for more equal pay. *Perspectives on Psychological Science*, 9(6):587–593, 2014.

[35] Gebhard Kirchgässner. Soft paternalism, merit goods, and normative individualism. *European Journal of Law and Economics*, 43(1):125–152, 2017.

[36] Saranga Komanduri, Richard Shay, Patrick Gage Kelley, Michelle L Mazurek, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, and Serge Egelman. Of passwords and people: Measuring the effect of password-composition policies. In *SIGCHI Conference on Human Factors in Computing Systems*, pages 2595–2604, 2011.

[37] Botond Kőszegi and Matthew Rabin. Reference-dependent risk attitudes. *American Economic Review*, 97(4):1047–1073, 2007.

[38] Botond Kőszegi and Matthew Rabin. Reference-dependent consumption plans. *American Economic Review*, 99(3):909–36, 2009.

[39] Steven D. Levitt, John A. List, Susanne Neckermann, and Sally Sadoff. The behavioralist goes to school: Leveraging behavioral economics to improve educational performance. *American Economic Journal: Economic Policy*, 8(4):183–219, 2016.

[40] Guocheng Liao, Xu Chen, and Jianwei Huang. Optimal privacy-preserving data collection: A prospect theory perspective. In *IEEE Global Communications Conference*, pages 1–6, 2017.

[41] Guocheng Liao, Xu Chen, and Jianwei Huang. Prospect theoretic analysis of privacy-preserving mechanism. *Transactions on Networking*, 28(1):71–83, 2019.

[42] Yiran Ma and Eleanor Birrell. Prospective consent: The effect of framing on cookie consent decisions. In *SIGCHI Conference on Human Factors in Computing Systems Extended Abstracts*, pages 1–6, 2022.

[43] Michelle L Mazurek, Saranga Komanduri, Timothy Vidas, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, Patrick Gage Kelley, Richard Shay, and Blase Ur. Measuring password guessability for an entire university. In *ACM Conference on Computer and Communications Security*, pages 173–186, 2013.

[44] Juanjuan Meng and Xi Weng. Can prospect theory explain the disposition effect? A new perspective on reference points. *Management Science*, 64(7):3331–3351, 2018.

[45] NordPass. Top 200 most common passwords of the year 2020. https://nordpass.com/most-common-passwords-list/.

[46] Michael I. Norton and Dan Ariely. Building a better america—one wealth quintile at a time. *Perspectives on psychological science*, 6(1):9–12, 2011.

[47] Michael I. Norton, David T. Neal, Cassandra L. Govan, Dan Ariely, and Elise Holland. The not-so-common-wealth of Australia: Evidence for a cross-cultural desire

for a more equal distribution of wealth. *Analyses of Social Issues and Public Policy*, 2014.

[48] Leilei Qu, Cheng Wang, Ruojin Xiao, Jianwei Hou, Wenchang Shi, and Bin Liang. Towards better security decisions: Applying prospect theory to cybersecurity. In *SIGCHI Conference on Human Factors in Computing Systems Extended Abstracts*, pages 1–6, 2019.

[49] Elissa M. Redmiles, Sean Kross, and Michelle L. Mazurek. How well do my results generalize? comparing security and privacy survey results from mturk, web, and telephone samples. In *IEEE Symposium on Security and Privacy*, pages 1326–1343, 2019.

[50] Anibal Sanjab, Walid Saad, and Tamer Başar. Prospect theory for enhanced cyber-physical security of drone delivery systems: A network interdiction game. In *IEEE International Conference on Communications*, pages 1–6, 2017.

[51] Agata Sawicka and Jose J. Gonzalez. Choice under risk in IT-environments according to cumulative prospect theory. In *21st International Conference of the System Dynamics Society, New York*, 2003.

[52] Jan Schnellenbach. Nudges and norms: On the political economy of soft paternalism. *European Journal of Political Economy*, 28(2):266–277, 2012.

[53] Neil J. Schroeder. Using prospect theory to investigate decision-making bias within an information security context. Technical report, Air Force Institution of Technology Wright-Patterson School of Engineering and Management, 2005.

[54] Hersh Shefrin and Meir Statman. The disposition to sell winners too early and ride losers too long: Theory and evidence. *The Journal of Finance*, 40(3):777–790, 1985.

[55] Erik Snowberg and Justin Wolfers. Explaining the favorite–long shot bias: Is it risk-love or misperceptions? *Journal of Political Economy*, 118(4):723–746, 2010.

[56] Justin Sydnor. (Over) insuring modest risks. *American Economic Journal: Applied Economics*, 2(4):177–99, 2010.

[57] Jenny Tang, Eleanor Birrell, and Ada Lerner. Replication: How well do my results generalize now? The external validity of online privacy and security surveys. In *Symposium on Usable Privacy and Security*, pages 367–385, 2022.

[58] Richard Thaler and Cass Sunstein. *Nudge: Improving Decisions About Health, Wealth, and Happiness*. Yale University Press, 2008.

[59] Richard H Thaler and Shlomo Benartzi. Save more tomorrow: Using behavioral economics to increase employee saving. *Journal of Political Economy*, 112(S1):S164–S187, 2004.

[60] Amos Tversky and Daniel Kahneman. The framing of decisions and the psychology of choice. *Science*, 211(4481):453–458, 1981.

[61] Amos Tversky and Daniel Kahneman. The framing of decisions and the evaluation of prospects. In *Studies in Logic and the Foundations of Mathematics*, volume 114, pages 503–520. Elsevier, 1986.

[62] Amos Tversky and Daniel Kahneman. Loss aversion in riskless choice: A reference-dependent model. *The Quarterly Journal of Economics*, 106(4):1039–1061, 1991.

[63] Amos Tversky and Daniel Kahneman. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5(4):297–323, 1992.

[64] Blase Ur, Patrick Gage Kelley, Saranga Komanduri, Joel Lee, Michael Maass, Michelle L. Mazurek, Timothy Passaro, Richard Shay, Timothy Vidas, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. How does your password measure up? The effect of strength meters on password creation. In *21st USENIX Security Symposium*, pages 65–80, 2012.

[65] Blase Ur, Fumiko Noma, Jonathan Bees, Sean M Segreti, Richard Shay, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. "I added '!'at the end to make it secure": Observing password creation in the lab. In *Eleventh Symposium On Usable Privacy and Security*, pages 123–140, 2015.

[66] Vilhelm Verendel. *A prospect theory approach to security*. Citeseer, 2008.

[67] John von Neumann and Oskar Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, 1944.

[68] Rick Wash. Folk models of home computer security. In *Proceedings of the Sixth Symposium on Usable Privacy and Security*, pages 1–16, 2010.

[69] Matt Weir, Sudhir Aggarwal, Michael Collins, and Henry Stern. Testing metrics for password creation policies by attacking large sets of revealed passwords. In *ACM Conference on Computer and Communications Security*, pages 162–175, 2010.

[70] Daniel Lowe Wheeler. zxcvbn: Low-budget password strength estimation. In *25th USENIX Security Symposium*, pages 157–173, 2016.

A Follow-up Survey Questions: User Study 1

1. Provide a brief description of the website you visited. (A few words or 1 sentence is sufficient.)

   [free response]

2. How strong was the password you chose when you created your account on the site?

   - Strong
   - Moderate
   - Weak

3. How much do you agree with the statement: A hacker would be likely to try to hack this site.

   - Completely agree
   - Somewhat agree
   - Neither agree nor disagree
   - Somewhat disagree
   - Completely disagree

4. How much do you agree with the statement: A hacker would be likely to successfully guess the password I used on this site.

   - Completely agree
   - Somewhat agree
   - Neither agree nor disagree
   - Somewhat disagree
   - Completely disagree

5. Is the password you used on this site a password that you also use on other sites?

   - Yes
   - No

6. How common are password stealing attacks?

   - Extremely common
   - Somewhat common
   - Neither common nor uncommon
   - Somewhat uncommon
   - Extremely uncommon

7. How could hackers potentially learn your password? Choose all that apply.

   - It is impossible for a hacker to learn my password.
   - If I accidentally download a virus, a malicious app, or a malicious attachment.
   - If I visit a sketchy or malicious website.
   - If I accidentally click on a phishing link and enter my credentials on a fake website.
   - If a hacker (or a program run by a hacker) guesses my password on the website.
   - If a hacker steals the files storing all passwords for the website.
   - Other: _____

8. How likely would it be for a password stealing attack to succeed if you use a weak password?

   - Extremely likely
   - Somewhat likely
   - Neither likely nor unlikely
   - Somewhat unlikely
   - Extremely unlikely

9. How likely would it be for a password stealing attack to succeed if you use a moderate password?

   - Extremely likely
   - Somewhat likely
   - Neither likely nor unlikely
   - Somewhat unlikely
   - Extremely unlikely

10. How likely would it be for a password stealing attack to succeed if you use a strong password?

    - Extremely likely
    - Somewhat likely
    - Neither likely nor unlikely
    - Somewhat unlikely
    - Extremely unlikely

11. Do you think upgrading your passwords can prevent password guessing?

    - Yes
    - Maybe
    - No

12. What could a hacker do if they successfully learn your password? Choose all that apply.

    - They could cause bugs (viruses can cause computers to crash, quit applications, erase important system files).
    - They could steal personal and financial information from individual computers, and send the information to criminal.
    - They could resell personal information.

- They could display annoying visual images on computers (a skull, advertising popups, or pornography).
- They could control the computer and use the computer to send information to others.
- They could use the computers to cause problems for third parties.
- Other: _____

13. Which of the following are likely to try to steal passwords? Choose all that apply.
    - A young computer geek who wants to show off or explore the internet
    - Criminals
    - Organizations and institutions
    - Other: _____
14. Which of the following best describes who is likely to be the target of a password stealing attack?
    - Hackers target everyone equally, and anyone is equally likely to have their password stolen
    - Hackers primarily target rich people
    - Hackers primarily target people with special privileges (e.g, system administrators)
    - Other: _____
15. What is your current age?
    - 18-24
    - 25-34
    - 35-44
    - 45-59
    - 60-74
    - 75+
16. What is your gender?
    - Man
    - Woman
    - Non-binary person
    - Other: _____
17. Choose one or more races that you consider yourself to be:
    - White
    - Black or African American
    - American Indian or Alaska Native
    - Asian
    - Pacific Islander or Native Hawaiian
    - Other: _____
18. Do you consider yourself to be Hispanic?
    - Yes
    - No

B Follow-up Survey Questions: User Study 2

1. Provide a brief description of the website you visited. (A few words or 1 sentence is sufficient.)

   [free response]

2. After you initially entered your Prolific ID and password, did you see a notice that looked like this (see picture above)?
    - Yes
    - No
    - I don't remember

   [Questions 3-8 were displayed only if participant answered yes to Question 2.]

3. In your opinion, how high would the risk of account compromise, financial loss, or identity theft be if you continued with your initial password?
    - Very high risk
    - High risk
    - Moderate risk
    - Low risk
    - Very low risk
4. In your opinion, how high would the risk of forgetting your password or losing access to the account be if you continued with your initial password?
    - Very high risk
    - High risk
    - Moderate risk
    - Low risk
    - Very low risk
5. In your opinion, how high would the risk of account compromise, financial loss, or identity theft be if you went back and chose a stronger password?
    - Very high risk
    - High risk
    - Moderate risk

- Low risk
- Very low risk

6. In your opinion, how high would the risk of forgetting your password or losing access to the account be if you went back and chose a stronger password?
   - Very high risk
   - High risk
   - Moderate risk
   - Low risk
   - Very low risk
7. Which choice was presented as the default option?
   - [Option to choose a stronger password. Exact wording depended on condition.]
   - [Option to continue with current password. Exact wording depended on condition.]
   - I don't remember.
8. Why did you choose that option?

   [free response]

   [Questions 9-16 were displayed only if participant answered selected the option to choose a stronger password for Question 7.]
9. In your opinion, how strong was the first password you chose?
   - Strong
   - Moderate
   - Weak
10. In your opinion, how likely is it that a hacker would be able to guess the first password you chose?
    - Very likely
    - Somewhat likely
    - Neither likely nor unlikely
    - Somewhat unlikely
    - Very unlikely
11. In your opinion, how likely is it that you would forget the first password you chose?
    - Very likely
    - Somewhat likely
    - Neither likely nor unlikely
    - Somewhat unlikely
    - Very unlikely
12. Is the first password you chose one that you use on another website or account?
    - Yes
    - No
    - No, but I use similar passwords for other websites or accounts
    - Prefer not to say
13. In your opinion, how strong was the second password you chose?
    - Strong
    - Moderate
    - Weak
14. In your opinion, how likely is it that a hacker would be able to guess the second password you chose?
    - Very likely
    - Somewhat likely
    - Neither likely nor unlikely
    - Somewhat unlikely
    - Very unlikely
15. In your opinion, how likely is it that you would forget the second password you chose?
    - Very likely
    - Somewhat likely
    - Neither likely nor unlikely
    - Somewhat unlikely
    - Very unlikely
16. Is the second password you chose one that you use on another website or account?
    - Yes
    - No
    - No, but I use similar passwords for other websites or accounts
    - Prefer not to say

    [Questions 17-20 were displayed if participant answered No to Question 2 or selected the option to choose a stronger password for Question 7.]
17. In your opinion, how strong was the password you chose?
    - Strong
    - Moderate
    - Weak

18. In your opinion, how likely is it that a hacker would be able to guess the password you chose?
    - Very likely
    - Somewhat likely
    - Neither likely nor unlikely
    - Somewhat unlikely
    - Very unlikely
19. In your opinion, how likely is it that you would forget the password you chose?
    - Very likely
    - Somewhat likely
    - Neither likely nor unlikely
    - Somewhat unlikely
    - Very unlikely
20. Is the password you chose one that you use on another website or account?
    - Yes
    - No
    - No, but I use similar passwords for other websites or accounts
    - Prefer not to say
21. How much do you agree with the statement: Having strong passwords is important to me.
    - Strongly disagree
    - Disagree
    - Neutral
    - Agree
    - Strongly agree
22. How much do you agree with the statement: Having passwords I will remember is important to me.
    - Strongly disagree
    - Disagree
    - Neutral
    - Agree
    - Strongly agree
23. In general, how likely is it that a hacker would be able to guess a strong password?
    - Very likely
    - Somewhat likely
    - Neither likely nor unlikely
    - Somewhat unlikely
    - very unlikely
24. In general, how likely is it that a hacker would be able to guess a moderate password?
    - Very likely
    - Somewhat likely
    - Neither likely nor unlikely
    - Somewhat unlikely
    - Very unlikely
25. In general, how likely is it that a hacker would be able to guess a weak password?
    - Very likely
    - Somewhat likely
    - Neither likely nor unlikely
    - Somewhat unlikely
    - Very unlikely
26. Which of the following best describes whose passwords a hacker would try to learn?
    - Hackers target everyone equally, and anyone is equally likely to have their password stolen
    - Hackers primarily target rich people
    - Hackers primarily target people with special privileges (e.g, system administrators)
    - Other: _____
27. What could a hacker potentially do if they successfully learn your password? Choose all that apply.
    - They could cause bugs (viruses can cause computers to crash, quit applications, erase important system files).
    - They could steal personal and financial information from individual computers, and send the information to criminal.
    - They could resell personal information.
    - They could display annoying visual images on computers (a skull, advertising popups, or pornography).
    - They could control the computer and use the computer to send information to others.
    - They could use the computers to cause problems for third parties.
    - Other: _____

[Participants were then asked to return to the website and log-in again.]

28. Were you able to successfully log-in to your account on the website?
    - Yes, I remembered my password
    - Yes, but it took me multiple tries to remember my password
    - No, I was unable to log-in to my account

    [Question 29 was only displayed if the participant said they were able to log-in to their account in Question 28.]

29. How did you remember your password for this account?
    - I wrote it down
    - I used a password manager
    - I have used this password before
    - I just remembered it
    - Other: _____

30. How much do you agree with the statement: I am proficient with the Internet and computers?
    - Strongly disagree
    - Disagree
    - Neutral
    - Agree
    - Strongly agree

31. How much do you agree with the statement: I am knowledgeable about security and privacy?
    - Strongly disagree
    - Disagree
    - Neutral
    - Agree
    - Strongly agree

32. What is your current age?
    - 18-24
    - 25-34
    - 35-44
    - 45-59
    - 60-74
    - 75+

33. What is your gender?
    - Man
    - Woman
    - Non-binary person
    - Prefer not to say
    - Prefer to self-describe: _____

34. Choose one or more races that you consider yourself to be:
    - White
    - Black or African American
    - American Indian or Alaska Native
    - Asian
    - Pacific Islander or Native Hawaiian
    - Other: _____

35. Do you consider yourself to be Hispanic/Latinx/Latine?
    - Yes
    - No

Who Comes Up with this Stuff? Interviewing Authors to Understand How They Produce Security Advice

Lorenzo Neil
North Carolina State University

Harshini Sri Ramulu
George Washington University

Yasemin Acar
Paderborn University &
George Washington University

Bradley Reaves
North Carolina State University

Abstract

Users have a wealth of available security advice — far too much, according to prior work. Experts and users alike struggle to prioritize and practice advised behaviours, negating both the advice's purpose and potentially their security. While the problem is clear, no rigorous studies have established the root causes of overproduction, lack of prioritization, or other problems with security advice. Without understanding the causes, we cannot hope to remedy their effects.

In this paper, we investigate the processes that authors follow to develop published security advice. In a semi-structured interview study with 21 advice writers, we asked about the authors' backgrounds, advice creation processes in their organizations, the parties involved, and how they decide to review, update, or publish new content. Among the 17 themes we identified from our interviews, we learned that authors seek to cover as much content as possible, leverage multiple diverse external sources for content, typically only review or update content after major security events, and make few if any conscious attempts to deprioritize or curate less essential content. We recommend that researchers develop methods for curating security advice and guidance on messaging for technically diverse user bases and that authors then judiciously identify key messaging ideas and schedule periodic proactive content reviews. If implemented, these actionable recommendations would help authors and users both reduce the burden of advice overproduction while improving compliance with secure computing practices.

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.
USENIX Symposium on Usable Privacy and Security (SOUPS) 2023.
August 6–8, 2023, Anaheim, CA, USA

1 Introduction

Most users of technology receive advice from experts to keep themselves and their devices safe. The overarching goal of security advice is to provide reliable and up-to-date security awareness and recommendations to end users so that they can practice secure behaviors. Employees of organizations may receive regular security advice and training from their employers, students may receive security advice from their schools and/or universities, and multiple government organizations including the US Department of State offer security advice to the general public [35]. "Second-hand" security advice abounds, proliferated by media outlets [17, 33], websites [22, 40, 44, 48], and peer users [36, 39].

End users are thus exposed to a sea of security advice and are unsure which advice is best suitable for them [38, 41, 44]. Experts also struggle to agree on which security advice should be prioritized. Previous related work demonstrated that experts list a total of 118 studied security behaviors as being the "Top 5" things users should do to protect themselves online [44]. A lack of consensus on the most important security imperatives leaves end users to themselves to prioritize and implement security advice. Authors who write security advice thus have to decide which advice is most important for their target audience, who already struggle to prioritize security advice.

In this paper, we seek to identify ground truth and root causes for why security advice varies in quality and prioritization by going to the source: authors themselves. We report findings from a semi-structured interview study with 21 authors of general security advice where we discussed the full process of advice creation from beginning to end. By "general security advice", we mean security advice just for the general public, or end users with a "general public level knowledge" of security. We investigated authors' backgrounds and motivations, asked about individual and organizational processes surrounding advice, and asked what they felt were the most challenging aspects of advice creation. In reporting the outcomes of these interviews, we present the following key

findings:

Content Creation Authors overwhelmingly perceive setting advice scope and technical level as a central challenge. Decisions about scope have major consequences on every other aspect of advice creation. Advice writers also report revising content when novel security threats or events prompt ad hoc additions. These challenges partially explain the overproduction and undercuration of security advice [44].

Internal and External Influences Authors reported input or oversight from a wide variety of stakeholders in their organizations, including technical and operational staff, legal departments, and even C-level executives. Legal regulations and technical standards also heavily influence advice content, with some organizations seeking comprehensive compliance or congruence with multiple sources. Authors consult a wide array of authoritative sources, with little consistency from author to author.

Recommendations Section 5.1 discusses implications of our findings and provides future recommendations. These recommendations include research on sound methodologies for curating and prioritizing advice, research establishing guidance on advice communication for users with varying levels of technical expertise, and for authors to proactively plan advice reviews to improve focus and not just augment advice.

2 Related Work

Security communications towards non-expert computer users comprise of more than just advice style communications. Informal sources of such communications consist of media [17, 26, 33], stories from peers [39, 50], and web pages containing computer security advice [22, 40, 48]. Formal sources of information that provide general security advice consist of security games [13, 47], nudges [4], training programs [24, 25, 25, 47, 51], and literature [29, 53]. The sources from which users retrieve general security advice impact their security mental models and then ultimately their decision making [7, 8, 33, 38, 41, 42]. Redmiles et al. suggest that user security decisions can be modeled as a function of past behavior and knowledge of costs, risks, and context of potential security decisions [43]. Understanding the prior work on how security knowledge is communicated to users, we look to investigate a specific but popular medium for end user security communications in security advice.

Since formal security guidance is not as widespread, online general security advice has been crafted to help users practice secure online habits. However, the general public is supplied with an overabundance of security advice and therefore has to prioritize which advice they will follow [21, 22, 33, 40, 42]. Prior work has suggested that users typically perform a cost-benefit analysis to determine if the benefits of the advice found are worth the cost of implementing the advice [6, 9, 10, 12, 15, 20, 21, 44, 45]. Herley et al. analyzed the cost-benefit tradeoff through various forms of security advice to determine much of the available security advice offers a poor cost-benefit tradeoff, therefore prompting users to reject advice [21].

The general public and technical experts have conflicting perceived responsibilities as to who is responsible for security advice implementations [19, 23, 52]. For example, Haney et al. found that smart home device users have perceived breakdowns in the relationship among who is responsible for the security of their devices between consumers, manufacturers, and relevant third parties such as the government [19]. We build from this prior work to investigate how advice writers determine their perceived set of responsibilities in writing security advice to their intended audience.

In recent years, many researchers have analyzed the quality of security advice for expert [2, 3, 18] and non-expert [5, 30, 31, 44] computer users. Prior work from Acar et al. evaluated the state of security practices from popular web resources that developers use for programming [2, 3]. They found a prevalence of security bugs within current guidance systems, therefore identifying insecure programming practices being advised to developers who seek these web resources [2, 3].

The work closest to ours is by Redmiles et al. [44], in which they investigate the quality of security and privacy advice on the web. Their work breaks down the quality evaluation by examining if security advice on the web is comprehensible, actionable, and effective. Their work concludes that the majority of the advice they investigated is perceived as actionable and comprehensible by both users and experts. However, users and experts both failed to come to a consensus as to what specific advice should be prioritized [44]. Not only did experts consider 89% of the 374 identified pieces of advice to be useful, they also struggled with internal consistency and alignment with the latest security guidelines.

The key challenges in addressing the volume and prioritization of security advice, as identified by Redmiles et al. [44], serve as motivations for our work. In this paper, we seek to understand what processes are implemented to write general security advice, as well as what decisions are considered when constructing the advice. We also seek to learn the challenges faced during the advice writing process that impacts the advice content. In doing so, we discover how advice writers gather information to draft advice content, how advice content is prioritized by the writers, and how procedural decision making and responsibilities are perceived by the writers.

3 Methods

We conducted 21 semi-structured interviews with authors of online security advice between September 2021 and March 2022 to understand the processes and decision-making that go into writing general security advice. We obtained written transcripts of audio interview recordings and analyzed the transcripts through deductive and inductive coding. Written transcripts were de-identified by replacing personally identifi-

able information (participant names, organizational names) with pseudonyms such as F001 for freelance workers, I002 for industry workers, and U007 for university IT and security workers. Participants were informed of the research goals and how their information would be protected in our screening survey consent forms. Our study protocol was approved by our University's Institutional Review Board (IRB).

3.1 Participant Recruitment

This project focuses on a specific expert population: authors of general security advice. We define such authors as those with professional experience in drafting content for general security advice. We recruit participants through purposive sampling of those who qualify through various recruitment channels, namely personal and professional contacts, social media advertising, recruitment on the freelancer platform Upwork, and directly emailing those who manage university security advice websites. We first directly recruited qualifying personal and professional contacts, then we posted messages on professional social media (Twitter, LinkedIn, and industry mailing lists) to solicit potential participants. We also advertised our study on the popular freelancer website Upwork [49] to recruit freelancers with professional experience in writing general security advice. After every interview, we asked participants if they knew other individuals that might qualify for our interview study. Finally, we reached out to IT help desks and information security or technology departments from U.S. universities found through a top national universities rankings website [32]. Here, we contacted 109 universities that provided both general security advice on their website and an email contact to either their IT help desk, information technology department, or security department. We contacted all potential participants with a recruitment email, linking to our public website which presented a study overview, supplementary information, and a link to the screening survey consent form.

Once we identified a potential participant and they replied with interest, we sent them a screening survey and informed consent form through Qualtrics [37]. The screening survey described our research goals at a high level, asked for consent to be video and/or audio recorded for the interview, and requested basic demographic information from the participant [37]. The survey also acted as a qualifier to ensure that the participant had prior professional experience with writing general security advice. Our screening survey asked participants to report on their security experience, such as how long they had been writing security advice, for what companies, and how they learned to write advice. Once eligible participants filled out the screening survey, we scheduled a one-hour interview with them. Participants were compensated with $30 per half hour for their participation in the interviews.

We concluded recruitment when we reached theoretical saturation; i.e., we discovered that participant responses were not presenting new information beyond data we had already collected [46]. Of the 21 participants, 12 were freelance workers, 6 were university security department staff, and 3 were industry workers. 9 of the freelancer workers wrote general security advice for external organizations, similar to a consulting role. 3 of the freelancer workers and 1 of the industry workers wrote general security advice for entities within their own organization or subsidiary organizations. The remaining 2 industry workers and 6 university employees wrote general security advice for the public consumer associated with their networks. Participants of different roles held different levels of involvement for specifically prioritizing the advice content. Awareness experts are communication specialists who reported "translating" advice from security employees to the general public, where security experts, technical experts, and analysts reported researching, brainstorming, or reviewing the audience's environment to formulate ideas for content. Demographic information on gender, advice generation role, and the organization type is presented in Table 1.

Table 1: Participant Demographics.

Gender	
Men: 13 (61.9%)	Women: 8 (38.1%)
Target Audience	
University Members: 6 (28.6%)	External Consulting: 9 (42.9%)
Public Consumers: 2 (9.5%)	Internal Consulting: 4 (19.0%)
Advice Generation Role	
Analyst: 3 (14.3%)	Security Expert: 14 (66.7%)
Awareness Expert: 3 (14.3%)	Technical Expert: 1 (4.8%)
Organization	
University: 6 (28.6%)	Industry: 4 (19.0%)
Defense Organization: 1 (4.8%)	Internet Provider: 2 (9.5%)
Government Office: 1 (4.8%)	Security Provider: 7 (33.3%)
Participant Group	
Freelance Workers: 12 (57.1%)	Industry Workers: 3 (14.3%)
University IT/Sec.: 6 (28.6%)	

3.2 Instrument Creation

Our goal in this study was to investigate the processes, decision-making, and challenges that play a role in the creation of general security advice. Based on our research questions, we drafted an initial set of high level questions. Using this draft, we then conducted two practice interviews and one pilot interview; our pilot was with a researcher who has experience writing general security advice. Based on these interviews, we revised our interview guide into three background questions and nine high level questions corresponding to our research questions, each with sub-question-level prompts.

Our interview guide contains questions about processes, decision-making, and challenges, such as *"Can you tell me about how security advice gets made and distributed at your organization?"*, *"Are there particular areas that are prioritized or discussed more in depth within the general security advice?"*, and *"Are there any tasks completed during general security advice creation/revisions that are challenging or time*

consuming?" The high-level version of our interview guide can be found in Appendix 7; we provide the full version with prompts in our replication package [1].

3.3 Interview Process

Choosing semi-structured qualitative interviews as our research method allowed us to ask broad questions about advice writing and then follow up with more specific questions where appropriate. Once we met participants virtually to be interviewed, we confirmed that they had read and understood the consent form and began recording. We reminded participants of the options to skip questions or terminate the interview, and we gave them a choice of audio or video recording. All interviews were conducted and recorded remotely via Zoom. Recordings were backed up with Open Broadcaster Software (OBS) [34]. All interviews were conducted in English and lasted between 30 minutes to an hour.

3.4 Data Protection

We took multiple steps to protect participants' privacy and data security. First, all participants were pseudonymized. Once interviews were completed, we saved audio recordings of the interviews and had them transcribed by a GDPR-compliant transcription service. Within each transcript, we thoroughly removed all personally identifiable participant data such as names, organizations, and demographic information. We also did not request identifying, confidential, or private information about our participants or their employers in our interviews. We used end-to-end encrypted tools in all of our study communications and data storage components.

3.5 Data Analysis

Data analysis was performed through inductive and deductive qualitative coding. We use coding not as a means to an end, but as a strategy to make sense of our data [14]. Codebook creation, coding, and discussion of disagreements helped us understand the data, formulate the themes that we describe in our results, and describe advice creation. We created our qualitative codebook based on our research questions, then expanded it with additional codes that emerged through open-coding the transcripts. The codebook was iterated over through weekly discussions with the team and through discussing and resolving disagreements between the first and second coder. The high-level version of our codebook can be found in the Appendix 8. The detailed operationalized codebook is included in our replication package. During the codebook development process, the coders independently double-coded 5 transcripts, with good inter-rater reliability at Krippendorff's alpha $>$ 0.75, and resolved all conflicts through discussion [16], after which the primary coder coded the remaining transcripts. With Krippendorff's alpha $>$ 0.75 for all transcripts, we are confident that our codebook is stable, represents our data well, and that our coding strategy was sound [28]. Altogether, the coders coded 21 and 5 transcripts, respectively.

3.6 Limitations

As with any interview study or self-reporting study, participant responses may be biased (e.g., self-reporting bias, social-desirability bias) or incomplete [27]. Specifically, some participants were not able to answer all questions we asked due to either a lack of access to that knowledge or a lack of experience.

Over half of our participants were freelancers (57.1%) who all reported writing advice for company employees in some consultation role. We also do not have detailed data on the audiences beyond what authors reported, though we feel it reasonable to assume only relatively large organizations have employees dedicated to this task. Some freelancers specialize in security advice, while others work on technical writing more broadly. Authors wrote for employees, university students, customers, or users, but in all cases, the authors assumed readers have a "general public level knowledge."

In theory, it is possible that paid participants would fraudulently participate in interviews. However, for participants recruited from advice websites, we are reasonably certain that they were genuine. For UpWork recruits, we specifically reached out to those who listed relevant expertise on their resumes; since writing is not UpWork's main focus, there is little incentive to fraudulently report this expertise. Participant pre-survey data and interview behavior also matched up. We are therefore reasonably sure that our participants were genuine.

Lastly, any study involving qualitative coding is subject to author biases and different coding strategies among coders. We address these biases in our investigation by first establishing a list of high level coding categories a priori that represented the high level questions that we developed in the creation of the interview, as mentioned in Section 3.2. A second research team member double coded five of the transcripts to ensure that the codebook was able to capture data that reflected our research questions, regardless of the coder.

4 Results

In this section, we present our qualitative findings from analyzing the in-depth perspectives of advice writers during general security advice creation. We use participant quotes to represent in their own words how participants answer our interview questions, and ultimately our research questions. We present exploratory findings for understanding the processes, decision making, and challenges encountered by the authors who write general security advice. Such volunteered information includes the company, target audience, and advice

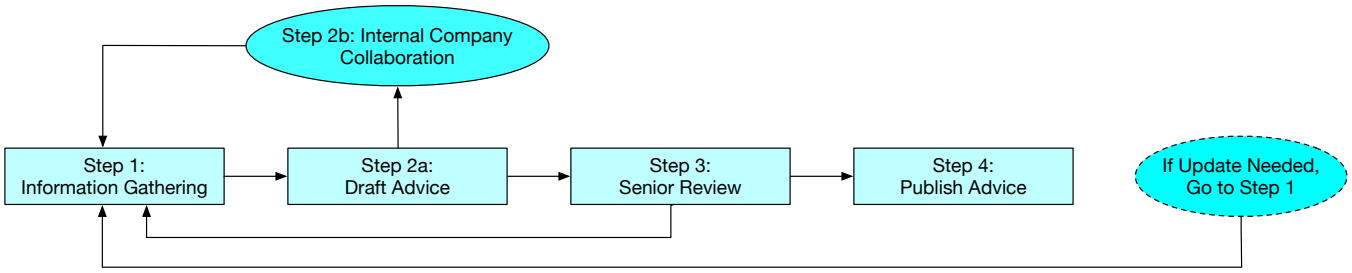


Figure 1: How experts write general security advice.

generation role they assumed while writing general security advice.

We find that participants followed a common advice creation for advice writing, found in Figure 1. This four-step process reflects advice-writers gathering information, drafting advice, sending the advice for review, and then publishing the advice, with options for iteration and further information gathering and collaboration in each stage. We also find that authors primarily prioritize and revise their advice content based either on current security trends or in response to security incidents. We organize the remainder of our results around this process. In Section 4.1, we describe how participants gather information for their advice content. Next is Section 4.2, where we explain the decision making that advice authors make when drafting and revising the advice. Section 4.3 details how advice authors collaborate with different internal company departments on the drafted advice before it is reviewed by senior-level employees and the company's legal department, in the case the legal department was involved. Lastly, Section 4.4 reports challenges participants mentioned for writing general security advice and also what improvements they stated could help general security advice writing.

4.1 Information Gathering

Here we explore findings from the first phase of advice writing, and information gathering. In this phase, authors are simultaneously gathering specific pieces of information (e.g., "prefer long passphrases") and establishing the scope of the advice more generally (e.g., "we should discuss password strength").

Theme 1: Advice writers research their environment to scope their advice. Some participants indicated that clients or stakeholders had already defined the scope of their advice documents. In the majority of cases, though, participants were left to figure out what advice was needed based on their own research of the client environment or what would be compliant with the organization. In these situations, authors develop a conceptual model that identifies what issues need to be addressed within the advice. This model is based on the organization, technology, and problems an organization currently faces (e.g. what areas where they are weak in security).Two broad approaches emerged from the interviews: holistic review and gap analysis. The primary difference between them is whether the advice is being written "from scratch" or to supplement existing advice.

As an example of holistic review, F020 explained they begin by determining "what are the devices, connected devices, the architecture and design of the network as well. How the company or how the devices are connected to each other and how the information is being transferred and sent between all the devices." From there, security advice is drafted for the elements of the system. In a gap analysis process, authors compare their architecture and operational environment to the advice available, and where advice is not present for an area that motivates additional material:

> "*Honestly, I report directly to the CISO. We meet every week for an hour at least, and we try to stay in lockstep about where the gaps are with what our incident response and governance compliance, and all those other teams are doing that can be addressed through getting educational materials out there in front of people.*" — U019

Theme 2: Advice writers base their own writing on multiple distinct external sources. 20 out of 21 participants stated they refer to an external source for sample information to include in their advice. These external sources may be regulations, technical standards, or industry or government agency documentation. Participants reported using multiple types of source for their content. 12/21 participants refer to legal regulations, specifically privacy regulations like GDPR(7/21 participants) and HIPAA(4/21 participants). One participant said about regulations they refer to:

> "*I would even say GDPR, where it clearly mentioned 'portal', 'what is personal data', 'what kind of controls', 'how consent should look.' Because, so easily, every company after the Internet boom — every website — was collecting data randomly from users without their consent.*" — F008

Participants specifically mentioned reviewing ISO standards (10/21), NIST publications (9/21), and PCI-DSS (8/21). In

some cases, participants' employers may suggest that the advice they publish should comply with regulations or standards the company adheres to. In others, the participants rely on these documents primarily to influence content:

> "*Normally, most of the companies we experienced are starting, and for certain companies, we normally recommend starting with NIST and ISO; they are well known, and their resources are well known, and they are well used.*" — F006

Authors also seek out external sources because the target audience may need to understand how to use specific software or how to mitigate a specific security issue. Sample advice in online web postings from government agencies was also referenced by participants as additional guidance for the advice they write, as one participant states:

> "*I use also some additional resources, such as NCES.ED.gov. This is the link that I also found on some extra information about security agreements, about some templates, for example.*" — F020

In aggregate, our participants cited a total of *20 distinct external sources* for their own work. Overall, these sources tended to be authoritative, so the content is likely correct. However, if these "upstream" standards documents were not properly scoped or written solely for technical experts, those issues may propagate "downstream" to general advice.

Some sources (e.g., GDPR, NIST publications) saw wide usage by a plurality of participants. However, of the total 20 cited external sources, only 5 were cited by 6 or more of participants.

4.2 Advice Drafting and Decision-Making

In the second step, advice writers draft advice based on gathered information and specific decision-making. Decision-making that impacts the advice content includes considering what areas of advice to prioritize, perceived responsibility authors held in advice writing, and addressing the usability of the advice.

Theme 3: Advice writers prioritize content in response to specific incidents or current trends. 13 participants indicated that specific security incidents or current industry trends were key prioritization factors. These participants form half of the participants in each of our four advice generation roles and form the majority of participants in each of the six organization groups. F006 describes how remote access became important during the COVID-19 pandemic:

> "*From my experience, most of the companies right now have employed remote work for their employees... They have a higher risk of having their data breached or compromised. So for me, what I would say, the remote access policy is the most important in this area, and we have to look into that because it's easy to compromise and very hard to detect what has happened.*" — F006

One participant described how incidents at other organizations led to advice creation:

> "*Sometimes if they give advice, it's based on something that's going on. For example, in a ransomware attack, they say, "Okay, universities are confronted with this type of attack, we should be careful with this.""* — I002

Advice creation may also correspond to events within the organization.

> "*I would say anything time critical is going to be prioritized, so if a change is happening and there's a deadline by which a user is going to need to make a change in accordance with whatever it is that's occurring, or people need to know about this change before it occurs, something like that is certainly going to be a priority.*" — U007

Theme 4: Advice writers most commonly cover online fraud and password security. While prioritizing advice on current trends or incidents was common, several specific areas were highlighted by participants. 10 participants indicated a priority on online fraud (e.g., phishing, social engineering, identity theft, email scams). On online fraud, U018 said:

> "*I've seen a lot of advice that we've been putting out regarding job scams or email scams. Phishing is a big one that we've put out a lot of general guidance on. Those are the only big ones that really come to mind is the job scams and the phishing.*" — U018

Five participants mentioned password security and authentication. On authentication, I002 said:

> "*And then, of course, password security is also a very important topic on everything that has to do with password security, like use of password managers, multi-factor authentication.*" — I002

Theme 5: Advice updates are reactive. 16 participants reported updating advice after new security trends or incidents become prominent. This theme of revising advice to reflect current trends or security incidents was also mentioned by at least half of participants within all six organization groups. A theme that indicates prioritizing content over novel or recurring advice topics. A majority of participants within the analyst, awareness expert, and security expert advice generation roles responded similarly. On reacting to new security incidents, U018 said:

> "*Any new information that we find we like to provide to the community so that they're aware of the new ways that these scammers or the phishers or the bad actors are trying to get to them.*" — U018

Theme 6: Advice writers cover a wide variety of less-common topics. Outside of online fraud and password security, participants mentioned prioritizing 12 other areas of advice, but no more than 4 participants mentioned any particular topic. Participants mentioned areas like remote access policy management, frameworks, organizational policies, incident response, and risk assessment. These trends concur with prior work that indicated a lack of consensus on advice prioritization [44].

A potential explanation may be that organizations have different advice needs, and our interviews support that interpretation. While fraud advice was the highest mentioned prioritized advice, it was only mentioned by half of the participants within only three different company types, respectively. Every participant representing either a defense company or internet provider mentioned prioritizing fraud advice, and five out of six total university workers mentioned prioritizing fraud advice. The rest of the advice areas are sparsely mentioned among the different participant sub-groups. This finding may also be caused by our Theme 5 findings in that new advice topics are constantly being addressed, given whatever security incident or trend is current.

Theme 7: Advice writers rarely curate advice by intentionally deprioritizing topics. Most participants did not provide significant responses when asked what areas of security advice were *not* prioritized or why. Four participants said advice for obsolete or deprecated technology would be removed or deprioritized. When asked, F016 replied: "*There's a lot of old depreciated functionality in Microsoft Windows.*" (F016). Others mentioned deprioritizing impractical or overly technical advice, though in at least one case this deprioritization was reluctant. For example, I003 mentioned encryption advice as a topic to deprioritize given being "too technical" for general security advice.

.

Theme 8: Advice writers consider usability, but without a consistent or systematic methodology. 20/21 participants mentioned an attempt to make their general security advice usable. However, participants mentioned 9 different methods to address usability, none of which were mentioned by more than 8 participants. 8 participants stated they simplify the technical language of the advice so that the general public can understand the advice. Participants also mentioned using visualizations and graphics to enhance usability:

> "*Our job is to translate this to something less technical and comprehensible for a broad public. Make it sometimes also a bit more visual — more attractive for users.*" — I002

Six participants determined if advice was usable by considering how they themselves would follow the advice.

> "*One of the things that we look at as we're writing the advice, how would we implement it? If we can't figure out how to implement our own recommendations within our own teams... how could we possibly expect people to be able to implement this?*" — I004.

Some participants considered usability but without a specific method for writing or evaluating advice usability.

> "*It's not really a process per se that is implemented. It's just something that we keep in mind to make it as user-friendly as possible.*"" — U018

Theme 9: Help desks are considered a backstop for unclear advice. 20 participants stated their organization had a team (such as a help desk) that could address users' security questions. U007 noted that they assumed that if advice is unclear, the help desk would correct the issue.

> "*Our expectation in every case is if the user is looking at instructional pages and they're confused about what they mean, or basically confused about anything that they see on the central IT website, that they would contact the help desk.*" — U007

We believe that this perspective may be optimistic about users' likelihood of asking questions before engaging in an unsafe action, and in any case this would be an interesting perspective for future work. On the other hand, if help desks receive the same questions frequently, it may be an indicator to authors that they should update advice on a topic.

Theme 10: Advice writers claim a wide range of responsibilities when writing general security advice. Prior work established a mismatch between users' and manufacturers' expectations of each other in smart homes [19]. Inspired by this effect, we asked participants about how much responsibility they or their company bear in advising users of secure practices. Overall, participants gave answers ranging from high levels of responsibility to virtually no responsibility. We also noticed that responses differed between the participants' roles, though we do not claim a literal correlation because this is an initial qualitative study. One of the university participants noted they take extreme ownership of their users' security education, and they stated their team's goal is to: "*help them, guide them, and educate them, and basically be a partner with them*" (U018). On the other hand, a freelancer gave a different perspective on perceived responsibility:

> "*Well, we give advice for the sake of advice. We want to be sure that we are not promoting what we do or represent. We are not trying to sell, like force you to patronize or do business with us*" — F011.

An industry worker mentioned assuming varying levels of responsibility depending on the content:

> "*It depends on what we're writing and why. If I'm writing the security advice or scoping the payment card industry data security standards audits, I'm just going to lean towards every time writing advice that would limit the scope of our environment because obviously, that makes it easier for us to pass the audit. If I am writing advice for, say, being ready to do something with data privacy, for instance, marking data as highly confidential things like that. I'm going to be as broad as I can be because I want everything protected*" — I004.

Overall, we recorded seven different categories of responsibility suggested by participants, though none were mentioned by more than six participants. Other levels of perceived responsibility included writing advice to comply with standards, motivating changes in security behavior, or going beyond documents to offer security workshops.

4.3 Collaboration and Review

In step two of the writing process, advice writers collaborate with internal company departments who review the advice. Then, in step three, the advice is sent to senior-level employees for final review and approval. In both steps, advice writers revise until the requested revisions are approved by the relevant stakeholder. Otherwise, the advice is approved and then published in step four.

Theme 11: Advice writing is distributed and collaborative. 16 participants stated there were multiple writers who worked on content, and they stressed the importance of multiple writers to lessen workload and include multiple perspectives:

> "*Currently, I tend to get most of the cybersecurity writing assignments in our group. I wouldn't necessarily say all, because we do have a focus on trying to develop bench strength, and within our group there are also people who are responsible for the communication regarding specific IT services offered by departments.*" — U010

18 participants mentioned collaborating with internal company departments. Security (12/21), marketing and communications (7/21), and human resources (HR) (3/21) were the most mentioned departments for advice collaboration. Participants noted that the mix of security experts and non-experts helped create content that is technically accurate and understandable to a broad audience:

> "*The central IT communications group (none of them have been members of the information security group) [works] closely with information security. So if they're writing about something that's a security issue, they're going to be corresponding with one or more people within information security making sure that they have the details right in their write-up. Or they may start with a couple of paragraphs that someone in information security supplied, and then they'll write their page around that to make sure that they're getting the technical details right.*" — U007

Theme 12: Advice is routinely reviewed by senior personnel. 17 participants mentioned submitting proposed general security advice to senior level employees (management or advisory board) for review, feedback, and approval. The university employees and industry workers we interviewed keep their advice within their own group before it is sent to their own management who then approves the advice. Freelancers writing for other organizations submit their advice to the management of their client company for review and approval. For example, participant F014 mentioned review by C-level executives.

Theme 13: In-house legal counsel can be heavily involved in advice creation. We reasoned at the beginning of the study that one cause of the overall lack of prioritization of security advice might be organizations writing comprehensively to limit liability rather than focus on the most likely issues. Therefore, we specifically ask about the involvement of counsel.

7 participants confirmed that their legal department was involved to ensure that the advice met certain legal standards and was up-to-date with the current law. As F011 explains:

> "*We believed that being a good lawyer does not mean being a good security expert, . . . so [lawyers] just review. If they feel something should be removed on legal grounds, they advise us. nd if they feel that we need to include some things based on legal grounds, they let us know. We include those things and then send that back to them for review, and then we go back and forth until they are satisfied with the documents.*" — F011

Another seven participants stated their legal was occasionally involved depending on the content or intended audience:

> " *If it comes to data protection specifically or a GDPR specifically, yes [legal is involved], because they would be the experts. For the rest, no.*" — I003.

The remaining seven participants indicated that they were either uncertain of the role of legal counsel or were certain that legal counsel was not involved. While legal counsel indeed influences content, we conclude that the extent of their influence does not explain the breadth and variety of security advice reported in the literature.

4.4 Reported Challenges and Improvements

Theme 14: Advice writers struggle to scope advice for broad audiences who lack fundamental security knowledge. 12 participants specifically mentioned difficulty in identifying not just the necessary topics for their intended audience, but also explaining relevant solutions for diverse technical settings. Simplifying advice content was recorded in responses by participants who mentioned it as a possible method to improve future general security advice. Throughout the study, participants gave examples of how both the intended audiences and teammates they worked with during advice writing came from different backgrounds and therefore all have different levels of security awareness. Therefore, it is important that general security advice be simplified for all intended audiences:

> "*I think not underestimating your audience, trying to empathize, trying to put things in terms where they understand that what we're helping them with is really the thing that they want the most.*" — U019

This challenge is especially seen in advice for employees who need to use software but lack necessary security awareness. Specifically, writing advice content that accounts for accurate assumptions about the intended user's security knowledge. One participant stated it is easier to advise more experienced clients who have security knowledge than those who are less experienced. Another participant mentioned that

> "*The problem internally was I think mostly that our IT teams supposed that everyone knew that we had a password manager, and they knew how to use it. But basically, this wasn't the case because a lot of people — I think also a lot of new people working for the organization — didn't know that it existed*" — I002

Theme 15: Advice writers value direct security training, despite its costs. Participants perceived that rectifying gaps in fundamental security knowledge is difficult:

> "*The most time-consuming task is when a company needs to train their employees, which takes a little time to train and give awareness to the employees.*" — F006

They also still recommend direct training.

> "*We believe in constant improvement. We believe in trainings. We believe in our teleconferences. We believe in individual investments in their whole training. So we encourage our team members to learn more. As a company, we try to find where we can get the kinds of trainings, conferences, events and all of that so that we can get updated on the current trends and security.*" — F011

One participant noted an alternative approach to traditional trainings:

> "*During the Cybersecurity Awareness Month we have games (particularly online, given the COVID). I'm astounded at how many people reach out and want to play these games for $25–$50 gift cards. So provide more games to attract people and use that to ask them questions. Perhaps one game session focuses on multi-factor, and another game focuses on strong passwords.* " — U009

Theme 16: Participants recognize the need for proactive updates. Participants also desired to revise or audit published advice on a regular basis, as opposed to strictly reactively as discussed earlier.

> "*Somebody needs to be looking through the pages and making sure that they're still relevant and that they still have current information, and I think that's an area that we're not really that good at.*" — U007

Theme 17: Collaboration and review leads to delay in publishing advice. We previously discussed how authors credited collaboration and review with leading to more readable and appropriate advice. Participants mentioned that collaborators who either are not consistent in their practices or do not meet important guidelines when writing the advice are challenging to work with. 6 participants mentioned that collaboration leads to delay because it requires the time of multiple busy parties. Time from senior stakeholders is even more difficult:

> "*A lot of times, those stakeholders are upper management. It's hard to get on their schedules. So there's always room for improvement in that. Sometimes the process takes too long because you're literally waiting for a day when you can get four people in a room together, and it's two weeks out. So there would be room for improvement there. Maybe that's top-down buy-in where they say this is more important than anything else; make time for it, which only happens after a breach. I would say those are two areas that would make things easier.*" — I004

5 Discussion

In this section, we contextualize our results with prior literature on advice prioritization, procedural decision making, and perceived responsibility in end user security. We then discuss how these findings can promote better practices for curating general security advice with methodological recommendations for both advice writers and their organizations.

5.1 Identifying Lack of Consensus

Advice Prioritization Participants prioritized their advice to reflect current security trends or respond to security incidents during advice construction and revision. Prioritized advice topics experience variable attention over time and may undergo fluctuating cycles of prioritization. Similarly, advice revisions exhibit a reactive manner in an attempt to constantly keep up with the latest security incidents and trends. While it may seem appropriate to prioritize content like this, it leads to a possible overproduction of advice on numerous security topics. We see this in Theme 6 as participants mention a total of 14 topics they prioritize in their advice writing.

Differences in prioritized advice topics among our study population may also be likely contributed by the differences in specific target audiences, roles, and organizations. However, even participants among the same groups sparsely agreed on which specific topics of advice they should prioritize. Rather, they instead looked to whatever novel security threat they determined they needed to cover. Relying on the latest threats and trends for advice writing also makes it more difficult to determine which security advice should *not* be prioritized. Outside of not covering obsolete or impractical advice, advice writers rarely provided significant responses to how they would deprioritize security advice.

We believe the reluctance to curate or deprioritize content partially explains the advice prioritization crisis documented in prior work. To borrow a common expression, "if everything is a priority, nothing is." Practitioners recognize that the attack surface for modern computing is vast and ever-changing. However, they may fail to account for the effort and opportunity cost to users caused by comprehensive advice. Our findings on the curation of security advice from advice writers add context to a continued theme from prior work indicating a lack of consensus on which general security advice should be prioritized [44]. Our results show that content covered in security advice is not curated to cover perennial topics, rather it is curated to cover many novel topics. Focusing on novel threats instead of perennial threats for security advice may contribute to overwhelming end-users with security advice they do not need. This implies that much general security advice found online may either be outdated or less relevant than when it was first written. Users of varying levels of security experience are left on their own to distinguish which security advice they actually need or is still important. While security experts are better equipped to make these important choices, end users lacking proper security awareness are less likely to understand the distinction between outdated and relevant security advice. This increases the number of security topics that end-users have to read through in advice and determine which advice they should prioritize. We make recommendations for a more proactive approach to advice and improving the lack of consensus for prioritizing general security advice in Section 5.2.

Procedural Decision Making A lack of consensus in procedural decision making was also prevalent in our findings. This is first observed among our Theme 1 and 2 findings which describe how information is gathered for their advice. We discovered that advice writers experience challenges in identifying key aspects of the scope of the advice they are tasked to write for their intended audience. This is a critical challenge given that most participants in our study state that information gathering is their first task in writing advice and sets the foundation for the scope of the content. If advice writers then make decisions on prioritization given an insufficient foundation, their advice will experience variable prioritization, given their inability to consistently define a scope for their advice. P009 stated they would prefer their advice to "target messages to students. It's a challenge. What we create is available to them, but how it's delivered and where it's delivered should be better targeted." Participants among all of the recruited groups experienced this challenge regardless of their organization or intended audience. General security advice affects end users in universities, organizational employees, and many other types of audiences who lack adequate general security awareness. A lack of consistency in properly identifying how general security advice should be written leads to a lack of consensus in advice prioritization from experts, which then trickles down to the lack of advice prioritization among general users.

Advice writers also refer to multiple distinct sources of sample advice that include any legal regulations, technical standards, or other organizational entities. In total, we identified 20 different external sources that participants mentioned they use to influence their general security advice. Of the 20 different external sources mentioned, only four were mentioned by at least 30% of participants. Similar to how this work and prior work [44] demonstrate a lack of advice content prioritization, there is also a lack of consensus among organizations and advice writers on which external sources to pull sample advice from. We see this lack of consensus among participants of the same recruitment group and also between participants of different recruitment groups. Specifically, we observe differences in external sources cited for influencing advice content even among participants who share both the same intended audience and the type of organization they worked for. An upstream usage of distinct external sources and an inconsistent foundation for gathering information for advice add more clarity as to why security experts differ in security advice prioritization. As suggested in prior related work [19], we discuss the importance of both advice authors and organizations to formulate standards to consistently curate perennial topics for security advice in Section 5.2.

Lastly, it is evident there is no consensus on agreed-upon methods in which advice authors consider the usability of their general security advice. Also, it is unclear if participants are performing appropriate usability checks in their advice for their intended audience. This lack of consensus for methods

in considering the usability of general security advice construction may be due to the lack of experience or knowledge about usability from both the authors and organizations. Another possible factor could be the pressure to release advice for a client within a specific time frame, a challenge that is seen more often for industry or freelancer workers. This increased pressure to meet deadlines to produce content may come at the expense of thoroughly considering the usability of general security advice. Regardless, a lack of consensus on which usability methods to implement to write general security advice can create advice of variable degrees of usability. Prior work has shown that end users follow or reject advice based on perceived cost and benefit analyses of the advice [6, 10, 15, 21]. This adds another burden that end users have to consider when deciding what general security advice they should prioritize. On top of deciding if advice presented to them is relevant to their needs, they also have to consider if that advice is worth implementing given the opportunity costs of implementing the advice. We recommend that specific usability tactics should become standards agreed upon by experts and authors of general security advice in order to lessen this burden. We explain this in further detail in Section 5.2

Perceptions on Security Responsibilities Previous work by Haney et al. [19] identified an interdependent relationship within user perceptions of the responsibility of smart home device privacy and security between three actors, namely the smart home device end users themselves, device manufacturers, and third parties such as government or regulatory bodies. In that paper, it was reported that users based their actions on that perceived interdependent relationship and therefore would not consider themselves the sole protector of the security of their smart home devices. However, manufacturers and regulatory third parties do not always act on this interdependent relationship, thus there exist gaps in understanding of the responsibilities for each party. In our work, we discovered several different responses for how authors of general security advice assess their perceived responsibility in advice writing. Some participants only wrote advice to comply with standards or requirements their organization enforces whereas other participants emphasized a need to educate their intended audience to develop better security decision making habits. Overall, there is no agreed upon consensus for what levels of responsibility that authors of general security advice or their organizations should assume in assisting their target audience. Our findings support results from previous work and highlight gaps in responsibility assumed not just between the general public and entities providing content, but also between the parties responsible for generating general security advice for users. Gaps in perceptions of shared responsibilities among the experts and end users fall further than just for smart home users, but also for both the general public and organizations who seek assistance with general security. A lack of agreed upon consensus on perceived responsibilities is apparent within both end users and the experts. This leads to increased confusion about which parties should be responsible for mitigating what issues or making security decisions for end-user software and technology. In Section 5.2, we add onto recommendations from previous related work [19] and make suggestions for explicit responsibility establishment for future general security advice creation.

5.2 Methodological Improvements for Advice Writing

We identify areas of improvement for general security advice construction by analyzing the current state of advice construction from the lens of the writers. We suggest methodological improvements for general security advice construction based on our current findings and findings from previous related work [11, 19, 44].

5.2.1 Develop the Domain of General Security Advising

A lack of consensus across multiple areas in the general security advice writing process indicates that its domain is not fully developed. We describe six domains of focus for professionals to consider when writing general security advice.

Resources/Technology: The biggest challenge participants mentioned when writing general security advice was defining the scope of advice content and how the advice can be broadly applied to all intended targeted audiences. P004 and P021 both advocated for the creation of an open source based repository to act as a research forum for advice writers. Both participants recommended that such a system should contain organized information about current general security questions or issues and that this system can be queried or allow open discussion between advice writers. Such a tool could be used by advice writers to both better discover what security topics need advice on and collaborate with other advice writers to come up with implementable solutions to communicate with their target audiences. While this is one idea of tool creation, future research may investigate the creation of new tools to help authors of general security advice identify advice content. Advice writers then would not have to rely on waiting for an incident to happen or a new trend to become popular in order to generate ideas for general security advice.

Relationships: Multiple parties are involved in discussing and reviewing general security advice before it is published. Understanding differences in viewpoints and requirements among organizational parties has been studied at a broader scope for internal corporate communications [11]. Critical challenges described in such work emphasized the importance of addressing communication related management problems between management and communications practitioners. Similarly, we recommend that advice writing parties each receive clear definitions on their responsibilities and roles during advice writing to reduce confusion among everyone involved.

Participants also mentioned that meeting with every party involved in the writing process can be time consuming since different parties (e.g., marketing, communications, security) have different schedules and duties to adhere to. Authors of general security advice should consider adopting a formal schedule and process that is agreed upon by all collaborating parties. This may help decrease the amount of time waiting for collaborating parties to meet and discuss advice content and how it should be implemented.

General Security Focus: Improving the security culture of both the intended audience and parties who write general security advice was recommended by participants as a means of improvement. Security awareness games, workshops, and other events to keep authors of general security updated on security trends help them stay connected to what issues are affecting their target audience. Providing the same programs for general users is also helpful for them to learn general security advice in a non-conventional way and should be considered by experts who write general security advice.

Content Improvement Metrics: Agreeing to a set of usability practices to make general security advice more usable can consist of the following: advice visualizations (e.g., diagrams, images, media, etc), and simplifications of overly technical words or phrases or templates to organize the advice content.

Community Support for Advice Writing: Establishing a community for advice writers to collaborate on ways to address common security threats through advice would greatly help writers in earlier stages of advice writing. Such a community can communicate by sharing best practices, sample advice from experts, or even recommendations for non advice-based approaches (e.g. games, workshops). Methods to evaluate the effectiveness of published advice over time can also be agreed upon by a community of advice writers. Academic researchers should also be involved in community collaboration in advice writing.

Human-Centered Engagement in Earlier Writing Stages: Advice writers in our study rarely mentioned user interaction with their intended audience as a means to generate content. Earlier stages in the writing process would benefit greatly from engaging directly with their intended audience to learn about security problems they may encounter. These direct interactions can help inform writers on which topics to prioritize, as well as how understandable or actionable their advice is. Writers should also gauge their users on whether the volume of advice is too high.

5.2.2 Proactive Advice Updates and Curation

Proactive Advice Updates: Implementing a proactive manner of updating or reviewing general security advice better ensures the advice is up to date. Participants mentioned performing more frequent check ups or audits of general security advice helps maintain the relevancy of the advice. Without consistent content audits, there increases the chances of advice becoming stagnant or not reflective of the current environment. Therefore, adopting a proactive approach to reviewing general security advice on a timely basis prevents the presence of outdated advice.

Establish a Set of Agreed Upon Standards for Advice Curation: We advocate there be a consistent standard to determine perennial areas of general security advice to cover. We say a set of standards since no one standard can be broadly applied to all advice of all intended audiences. Therefore, we suggest advice authors and industries communicate both what advice should be perennial and what advice should not be prioritized. Also, we suggest the research community investigate further what security advice end-users actually claim they need and if they are receiving that advice now.

6 Conclusion

In a semi-structured interview study with 21 authors of general security advice, we analyze the processes, decision making, and challenges that experts face when writing general security advice. We corroborate the lack of consensus on security advice as well as responsibility assignment from prior work. Our contribution gives insights into the context and reasons for this lack of consensus: advice writers struggle to define the advice scope and prioritize the information necessary to write advice, and must prioritize time-sensitive events over curating perennial advice. Based on our findings, we provide recommendations for how general security advice authors can better develop the domain of general security advice writing, implement proactive approaches towards writing and revising general security advice, and establish a set of agreed-upon standards for advice writers to reference when curating general security to end users. Addressing the lack of agreement on how general security should be advised from both the end user and advice writer side may improve the relationship between both parties in general security advice prioritization and perceived responsibilities.

References

[1] Anonymized replication package. https://advice22.netlify.app/.

[2] Yasemin Acar, Michael Backes, Sascha Fahl, Doowon Kim, Michelle L Mazurek, and Christian Stransky. You get where you're looking for: The impact of information sources on code security. In *2016 IEEE Symposium on Security and Privacy (SP)*, pages 289–305. IEEE, 2016.

[3] Yasemin Acar, Christian Stransky, Dominik Wermke, Charles Weir, Michelle L Mazurek, and Sascha Fahl. Developers need support, too: A survey of security advice for software developers. In *2017 IEEE Cybersecurity Development (SecDev)*, pages 22–26. IEEE, 2017.

[4] Alessandro Acquisti, Idris Adjerid, Rebecca Balebako, Laura Brandimarte, Lorrie Faith Cranor, Saranga Komanduri, Pedro Giovanni Leon, Norman Sadeh, Florian Schaub, Manya Sleeper, et al. Nudges for privacy and security: Understanding and assisting users' choices online. *ACM Computing Surveys (CSUR)*, 50(3):1–41, 2017.

[5] Devdatta Akhawe and Adrienne Porter Felt. Alice in warningland: A large-scale field study of browser security warning effectiveness. In *22nd USENIX Security Symposium (USENIX Security 13)*, pages 257–272, Washington, D.C., August 2013. USENIX Association.

[6] Adam Beautement, M Angela Sasse, and Mike Wonham. The compliance budget: managing security behaviour in organisations. In *Proceedings of the 2008 New Security Paradigms Workshop*, pages 47–58, 2008.

[7] Maia J Boyd, Jamar L Sullivan Jr, Marshini Chetty, and Blase Ur. Understanding the security and privacy advice given to black lives matter protesters. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2021.

[8] Cristian Bravo-Lillo, Lorrie Faith Cranor, Julie Downs, and Saranga Komanduri. Bridging the gap in computer security warnings: A mental model approach. *IEEE Security & Privacy*, 9(2):18–26, 2010.

[9] Karoline Busse, Julia Schäfer, and Matthew Smith. Replication: No one can hack my mind revisiting a study on expert and {Non-Expert} security practices and advice. In *Fifteenth Symposium on Usable Privacy and Security (SOUPS 2019)*, pages 117–136, 2019.

[10] Nicolas Christin, Serge Egelman, Timothy Vidas, and Jens Grossklags. It's all about the benjamins: An empirical study on incentivizing users to ignore security advice. In *International Conference on Financial Cryptography and Data Security*, pages 16–30. Springer, 2011.

[11] Joep P Cornelissen. Corporate communication: A guide to theory and practice. *Corporate Communication*, pages 1–336, 2020.

[12] Duy Dang-Pham, Siddhi Pittayachawan, and Vince Bruno. Why employees share information security advice? exploring the contributing factors and structural patterns of security advice sharing in the workplace. *Computers in Human Behavior*, 67:196–206, 2017.

[13] Tamara Denning, Adam Lerner, Adam Shostack, and Tadayoshi Kohno. Control-alt-hack: the design and evaluation of a card game for computer security awareness and education. In *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*, pages 915–928, 2013.

[14] Victoria Elliott. Thinking about the coding process in qualitative data analysis. *The Qualitative Report*, 23(11):2850–2861, 2018.

[15] Michael Fagan and Mohammad Maifi Hasan Khan. Why do they do what they do?: A study of what motivates users to (not) follow computer security advice. In *Twelfth symposium on usable privacy and security (SOUPS 2016)*, pages 59–75, 2016.

[16] Deen Freelon. *ReCal2: Reliability for 2 Coders*.

[17] Kelsey R. Fulton, Rebecca Gelles, Alexandra McKay, Yasmin Abdi, Richard Roberts, and Michelle L. Mazurek. The effect of entertainment media on mental models of computer security. In *Fifteenth Symposium on Usable Privacy and Security (SOUPS 2019)*, pages 79–95, Santa Clara, CA, August 2019. USENIX Association.

[18] Peter Leo Gorski, Luigi Lo Iacono, Dominik Wermke, Christian Stransky, Sebastian Möller, Yasemin Acar, and Sascha Fahl. Developers deserve security warnings, too: On the effect of integrated security advice on cryptographic {API} misuse. In *Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018)*, pages 265–281, 2018.

[19] Julie Haney, Yasemin Acar, and Susanne Furman. " it's the company, the government, you and i": User perceptions of responsibility for smart home privacy and security. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 411–428, 2021.

[20] Julie M Haney, Mary Theofanos, Yasemin Acar, and Sandra Spickard Prettyman. " we make it a big deal in the company": Security mindsets in organizations that develop cryptographic products. In *Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018)*, pages 357–373, 2018.

[21] Cormac Herley. So long, and no thanks for the externalities: the rational rejection of security advice by users. In *Proceedings of the 2009 Workshop on New Security Paradigms Workshop*, pages 133–144, 2009.

[22] Cormac Herley. More is not the answer. *IEEE Security & Privacy*, 12(1):14–19, 2013.

[23] Sri Lakshmi Kanniah and Mohd Naz'ri Mahrin. A review on factors influencing implementation of secure software development practices. *International Journal of Computer and Systems Engineering*, 10(8):3032–3039, 2016.

[24] Ponnurangam Kumaraguru. *Phishguru: a system for educating users about semantic attacks*. Carnegie Mellon University, 2009.

[25] Ponnurangam Kumaraguru, Justin Cranshaw, Alessandro Acquisti, Lorrie Cranor, Jason Hong, Mary Ann Blair, and Theodore Pham. School of phish: a real-world evaluation of anti-phishing training. In *Proceedings of the 5th Symposium on Usable Privacy and Security*, pages 1–12, 2009.

[26] Ponnurangam Kumaraguru, Steve Sheng, Alessandro Acquisti, Lorrie Faith Cranor, and Jason Hong. Teaching johnny not to fall for phish. *ACM Transactions on Internet Technology (TOIT)*, 10(2):1–31, 2010.

[27] Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser. *Research methods in human-computer interaction*. Morgan Kaufmann, 2017.

[28] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for cscw and hci practice. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–23, 2019.

[29] Christine Mekhail, Leah Zhang-Kennedy, and Sonia Chiasson. Visualizations to teach about mobile online privacy. In *Persuasive Technology Conference (poster)*, 2014.

[30] Lorenzo Neil, Yasemin Acar, and Bradley Reaves. Investigating Web Service Account Remediation Advice. In *Who Are You?! Adventures in Authentication Workshop*, WAY '20, pages 1–6, Virtual Conference, August 2020.

[31] Lorenzo Neil, Elijah Bouma-Sims, Evan Lafontaine, Yasemin Acar, and Bradley Reaves. Investigating web service account remediation advice. In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*, pages 359–376. USENIX Association, August 2021.

[32] U.S. News. Best national university rankings. https://www.usnews.com/best-colleges/rankings/national-universities?_mode=table. [Online; accessed February 23, 2022].

[33] James Nicholson, Lynne Coventry, and Pamela Briggs. " if it's important it will be a headline" cybersecurity information seeking in older adults. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–11, 2019.

[34] OBS. Open broadcaster software. https://obsproject.com/wiki/OBS-Studio-Overview. [Online; accessed February 23, 2022].

[35] U.S. Department of State. *Fraud Warning*. https://travel.state.gov/content/travel/en/us-visas.html/.

[36] Katharina Pfeffer, Alexandra Mai, Edgar Weippl, Emilee Rader, and Katharina Krombholz. Replication: Stories as informal lessons about security. In *Eighteenth Symposium on Usable Privacy and Security (SOUPS 2022)*, pages 1–18, 2022.

[37] Qualtrics. qualtrics. https://www.qualtrics.com/. [Online; accessed February 23, 2022].

[38] Emilee Rader and Rick Wash. Identifying patterns in informal sources of security information. *Journal of Cybersecurity*, 1(1):121–144, 2015.

[39] Emilee Rader, Rick Wash, and Brandon Brooks. Stories as informal lessons about security. In *Proceedings of the Eighth Symposium on Usable Privacy and Security*, pages 1–17, 2012.

[40] Elissa M Redmiles, Sean Kross, and Michelle L Mazurek. How i learned to be secure: a census-representative survey of security advice sources and behavior. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 666–677, 2016.

[41] Elissa M Redmiles, Sean Kross, and Michelle L Mazurek. Where is the digital divide? a survey of security, privacy, and socioeconomics. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 931–936, 2017.

[42] Elissa M Redmiles, Amelia R Malone, and Michelle L Mazurek. I think they're trying to tell me something: Advice sources and selection for digital security. In *2016 IEEE Symposium on Security and Privacy (SP)*, pages 272–288. IEEE, 2016.

[43] Elissa M Redmiles, Michelle L Mazurek, and John P Dickerson. Dancing pigs or externalities? measuring the rationality of security decisions. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pages 215–232, 2018.

[44] Elissa M Redmiles, Noel Warford, Amritha Jayanti, Aravind Koneru, Sean Kross, Miraida Morales, Rock Stevens, and Michelle L Mazurek. A comprehensive quality evaluation of security and privacy advice on the web. In *29th {USENIX} Security Symposium ({USENIX} Security 20)*, pages 89–108, 2020.

[45] Robert W Reeder, Iulia Ion, and Sunny Consolvo. 152 simple steps to stay safe online: Security advice for non-tech-savvy users. *IEEE Security & Privacy*, 15(5):55–64, 2017.

[46] Benjamin Saunders, Julius Sim, Tom Kingstone, Shula Baker, Jackie Waterfield, Bernadette Bartlam, Heather Burroughs, and Clare Jinks. Saturation in qualitative research: exploring its conceptualization and operationalization. *Quality & quantity*, 52(4):1893–1907, 2018.

[47] Steve Sheng, Bryant Magnien, Ponnurangam Kumaraguru, Alessandro Acquisti, Lorrie Faith Cranor, Jason Hong, and Elizabeth Nunge. Anti-phishing phil: the design and evaluation of a game that teaches people not to fall for phish. In *Proceedings of the 3rd symposium on Usable privacy and security*, pages 88–99, 2007.

[48] Sarah Turner, Jason Nurse, and Shujun Li. When googling it doesn't work: The challenge of finding security advice for smart home devices. In *International Symposium on Human Aspects of Information Security and Assurance*, pages 115–126. Springer, 2021.

[49] Upwork. Upwork. https://www.upwork.com/. [Online; accessed February 23, 2022].

[50] Rick Wash and Molly M Cooper. Who provides phishing training? facts, stories, and people like me. In *Proceedings of the 2018 chi conference on human factors in computing systems*, pages 1–12, 2018.

[51] Laurie Williams, Andrew Meneely, and Grant Shipley. Protection poker: The new software security" game". *IEEE Security & Privacy*, 8(3):14–20, 2010.

[52] Jing Xie, Heather Richter Lipford, and Bill Chu. Why do programmers make security errors? In *2011 IEEE symposium on visual languages and human-centric computing (VL/HCC)*, pages 161–164. IEEE, 2011.

[53] Leah Zhang-Kennedy, Sonia Chiasson, and Robert Biddle. The role of instructional design in persuasion: A comics approach for improving cybersecurity. *International Journal of Human-Computer Interaction*, 32(3):215–257, 2016.

7 Interview Guide

1. Ice Breaker Questions:
   (a) When and where did you learn to write general security advice?
   (b) What is your current occupation?
   (c) Can you describe the type of company you worked for when you wrote general security advice?
2. Can you tell me about how security advice gets made and distributed at your organization?
   (a) Is there a decision making model for the creation of security advice?
   (b) Are there any external sources used to provide sample advice that gets posted?
3. Can you tell me about the people or roles involved in the process?
   (a) Does a chain of command or hierarchy exist within the parties?
   (b) Are all of these parties involved with the company or external?
   (c) What is typically the experience or knowledge of parties in regards to computer security?"
4. Are there particular areas that are prioritized or discussed more in depth within the general security advice?
   (a) If so, what is the reason for this prioritization or focus into this area?
   (b) Are there any areas that are intentionally excluded from being covered in the security advice?
   (c) If so, what is the reason for not writing advice for this specific area?
5. Is the general security advice regularly updated or reviewed?
   (a) If so, what systems or procedures are in place to update/review the advice?
   (b) Were these systems/procedures always in place, or did an event or policy create them?
   (c) If possible to comment, are there legal practices or regulations that prompt the creation and/or regulation of the advice?
6. Is your company's legal department involved in the creation or even discussion of the general security advice?
   (a) If so and you are able to comment, are they able to edit or create any parts of the advice, or even recommend certain areas be covered?
7. If possible, can you comment on how much responsibility your organization claims in assisting in general security?
   (a) How much of the advice is well-meant, or meant to limit the reliability/responsibility of the service in security matters with general security?
8. Does your company have a team or group of individuals that handle general security internally?
   (a) Are they external workers?
   (b) Do they have expert experience in computer security?
9. When creating the general security advice, is there a thought process as to how actionable or practical the advice may be for the typical user?
10. These last questions are more so geared to your own experiences when creating the advice.
    (a) Are there any tasks completed during general security advice creation/revisions that are challenging or time consuming?
    (b) Have you ever thought about how general security advice for your company, or overall can be improved?

8 Codebook

High Level Codes	
Codes	**Code Explanations**
1a. Learn to write Advice	How the participant first learned to write general security advice.
1b. Occupational Role	Occupations for participants during the time they wrote general security advice.
1c. Companies	Places where the participant worked for and wrote the advice.
2a. Formal Writing Process	Any formal or structured process (Gap Analysis, SLA, defining scope, etc) used for writing advice.
2b. Informal Writing Process	Advice writing that is not dependent on any formal process. Rather, it is written in an informal or non-structured writing process.
2c. Legal or Non Legal Guidelines	Mandates, regulations, laws, or frameworks that were used to influence the advice. These are not solely or specifically technical, but apply to a wider range of compliance standards.
2d. Technical,Security Standards	Advice content is influenced by technical and/or security standards.
2e. External Entities	External entities (organization, group, company, etc) that authors seek for guidance on advice writing.
3a. Background,Experience	Backgrounds of fellow workers/teammates of advice authors.
3b. External Company Party Collaboration	Parties outside the primary advice construction group that collaborate in the advice writing process (outside or external to the company).
3c. Internal Company Party Collaboration	Parties outside the primary advice construction group that collaborate in the advice writing process (within the company).
3d. Writers	The number of people specified by the participant who helps physically write the advice.
4a. Most Prioritized Advice	Most common/prioritized topics of advice written.
4b. Least Prioritized Advice	Least common/prioritized topics of general security advice.
4c. Reasons Advice is Prioritized	Reasons or events that would cause the creation of general security advice.
4d. Reasons Advice is not Prioritized	Reasons certain advice has not been covered as much or prioritized.
5a. Revision Process	Processes and reasons to revise advice.
6. Company's legal department	Company's legal department involvement within the advice writing process.
7. Responsibilities	Responsibilities claimed by participant companies when creating the advice.
8. Internal Support	Support for clients that is internal or technical (not advice).
9. Advice Usability Thought Process	Though process or methods of improving actionability/usability of the advice.
10a. Challenges	Challenges with writing the advice.
10b. Improvements	Authors' opinions of how the advice writing process could be improved.

Towards Usable Security Analysis Tools for Trigger-Action Programming

McKenna McCall[1] *mckennak@cmu.edu*
Eric Zeng[1] *ericzeng@cmu.edu*
Faysal Hossain Shezan[2] *fs5ve@virginia.edu*
Mitchell Yang[1] *mfy@andrew.cmu.edu*
Lujo Bauer[1] *lbauer@cmu.edu*
Abhishek Bichhawat[3] *abhishek.b@iitgn.ac.in*
Camille Cobb[4] *camillec@illinois.edu*
Limin Jia[1] *liminjia@cmu.edu*
Yuan Tian[5] *yuant@ucla.edu*

[1]*Carnegie Mellon University* [2]*University of Virginia* [3]*IIT Gandhinagar* [4]*University of Illinois Urbana-Champaign* [5]*University of California, Los Angeles*

Abstract

Research has shown that trigger-action programming (TAP) is an intuitive way to automate smart home IoT devices, but can also lead to undesirable behaviors. For instance, if two TAP rules have the same trigger condition, but one locks a door while the other unlocks it, the user may believe the door is locked when it is not. Researchers have developed tools to identify buggy or undesirable TAP programs, but little work investigates the usability of the different user-interaction approaches implemented by the various tools.

This paper describes an exploratory study of the usability and utility of techniques proposed by TAP security analysis tools. We surveyed 447 Prolific users to evaluate their ability to write declarative policies, identify undesirable patterns in TAP rules (anti-patterns), and correct TAP program errors, as well as to understand whether proposed tools align with users' needs. We find considerable variation in participants' success rates writing policies and identifying anti-patterns. For some scenarios over 90% of participants wrote an appropriate policy, while for others nobody was successful. We also find that participants did not necessarily perceive the TAP anti-patterns flagged by tools as undesirable. Our work provides insight into real smart-home users' goals, highlights the importance of more rigorous evaluation of users' needs and usability issues when designing TAP security tools, and provides guidance to future tool development and TAP research.

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.
USENIX Symposium on Usable Privacy and Security (SOUPS) 2023.
August 6–8, 2023, Anaheim, CA, USA

1 Introduction

Platforms like IFTTT [13], SmartThings [20], and Home Assistant [10] allow users to create *home automations* to configure their smart homes. Home automations are often expressed as *trigger-action programming* (TAP) rules, which are an accessible way for people without programming experience to customize their smart-home devices [22]. A typical TAP rule format is: "IF *trigger* THEN *action*", where the *trigger* is an event that causes the *action*. For example, "IF Alice leaves home THEN lock door" locks the door whenever Alice leaves her home. More complex TAP rules can include conditional triggers or trigger multiple actions.

Users can accomplish more complex goals by writing multiple TAP rules, which we call TAP *programs*. When TAP rules are executed, they can generate events or alter the home environment, which may trigger other rules. For instance, consider the rules: "IF the user nears home THEN unlock door" and "IF door is unlocked THEN turn off security camera". Once the user reaches home, the first rule is triggered, and the resulting action (unlock door) triggers the second rule, causing the camera to stop recording.

Security and privacy risks of TAP Research has shown that users struggle to write complex TAP programs [12, 28] and reason about their behavior [12, 26], leading to problems ranging from safety risks like leaving doors unlocked [4, 5] to privacy risks like leaking sensitive data [3, 21]. Consider a situation where Bob has installed the rules "IF the time is 7PM THEN lock door" and "IF Bob arrives at home THEN unlock door". Here, Bob might mistakenly believe that the first rule will re-lock his door after he comes home at 8PM, a misunderstanding that could pose a safety risk. Further, even if Bob realizes there is a problem, he might have trouble finding a solution.

TAP security analysis tools To address these concerns, researchers have proposed tools to diagnose TAP program problems. Some tools detect *TAP anti-patterns*: recurring structures of TAP rules that can lead to unexpected or problematic behaviors [4, 5, 7, 16, 18, 24, 27]. Examples of anti-patterns

include rules that trigger conflicting behaviors in the same device, rules that might trigger each other (loops), and multiple rules triggering the same action.

Other tools verify that TAP programs adhere to *declarative policies* specified by users: the user specifies how they want their smart home to behave, (e.g., the front door should be locked at night), and the tool checks that no TAP rules violate the policy [2, 4, 5, 14, 15]. A proposed approach to specify policies are fill-in-the-blanks *policy templates* [16, 28]. For instance, Bob might write the policy "Door is unlocked should never be active while the time is after 7PM", where "_ should [always/never] be active while _" is the policy template.

However, there has been little work on whether these tools are easy to use and whether they satisfy users' actual needs. For example, it is not known which of the TAP anti-patterns identified by prior work are of concern to users, or which of the proposed templates for specifying declarative policies are easiest for users to understand and fill out correctly.

Research questions To gain insight into which TAP analysis tool user-interface approaches work well and which may need more refinement, we conducted an exploratory survey. We recruited 447 smart-home users from Prolific to study their perceptions of the problems that TAP security analysis tools address and the usability of the approaches proposed by these tools. In particular, we investigate the following questions:

- **RQ1**: What are smart-home users' motivations and goals for using TAP rules?
- **RQ2**: Can users successfully write *declarative policies* using the templates proposed in prior work? Which templates do people use most successfully, and which ones do people most prefer?
- **RQ3**: Given a TAP program, can users identify the *TAP anti-patterns* from prior work? Do users perceive anti-patterns as undesirable, and would they be interested in using a tool to find them?
- **RQ4**: What information should a tool provide that would be most helpful to aid users fixing policy violations or removing anti-patterns found in their TAP programs?

Overall, we show that smart-home users have the necessary skills to use TAP security analysis tools, but there is room for improvement. Participants were moderately successful at specifying declarative policies using templates, but struggled with some template formats. Participants also varied in their abilities to identify TAP anti-patterns. While some anti-patterns were viewed as desirable, participants generally agreed that it would be helpful to have a tool to identify them. Participants were most successful at repairing buggy TAP programs and completing partial programs that violated declarative policies when they were given some additional context about *the state of the home* when the policy violation would occur. Based on our results, some important shortcomings of these tools include confusingly worded policy templates and inconsistencies between user mental models of TAP rules and how they are modeled in the tools. Our findings provide guidance for both developers of TAP security analysis tools and researchers investigating the security and usability of TAP.

2 Background and Related Work

We broadly categorize TAP tools as ones that verify custom *declarative policies*, and ones that look for high-level *TAP anti-patterns*; several tools do both [4, 5, 16, 18, 27].

Tools with declarative policies Tools such as SIFT [15], Soteria [4], Salus [14], and AutoTap [28] typically require the user to specify which devices and TAP rules they have in their smart home, as well as how they want their devices to behave (a *policy*). The tool then checks the policy against a formal model of the user's home setup. Some tools perform this verification at run time [5], notifying users of potential violations as they interact with devices and trigger TAP rules.

The interfaces to specify policies vary in complexity between tools. Some tools require policies to be specified in a formal logic [4, 18, 27]. Others allow the user to write a policy as a series of conditions that should never happen [15] or via fill-in-the-blanks templates [16, 28], which are then translated to a more formal language. We focus on templates, as they require less training than writing policies in formal logic.

Chaining together smart devices through TAP programs leads to diverse and complex functionality, but also means that finding the source of a problem with a TAP program can be difficult [25]. Upon finding a policy violation, some tools supply a counterexample to help with debugging [4]. Others give hints about what led to the violation, like the rules responsible for the problem [16, 18]. In our study, we investigate what type of feedback is most helpful to the user trying to fix problems: identifying rules involved in the violation, supplying a full trace leading up to the violation, or something in between. Some tools go further and automatically synthesize fixes to suggest to the user [14, 28]. To narrow the scope of our study, we don't evaluate synthesis techniques but focus instead on what information tools can give users to help them understand the causes of violations and how to solve them. For tools that automatically fix problems, our results could still be relevant because users still need to specify a property (RQ2) and may benefit from additional feedback to help them understand what problem is being addressed (RQ4).

TAP anti-patterns Research has identified *TAP anti-patterns* [7, 11, 24] that can lead to unsafe or unpredictable behavior [1]. For instance, Bob's program from Section 1 exhibits an anti-pattern (*Opposite Behaviors* in Table 2) frequently identified in TAP research as problematic: if two rules trigger simultaneously with conflicting actions (door unlock and door lock) the result becomes unclear, which could lead to confusing or unsafe situations (such as Bob's door being unlocked when he believes it is locked). Other research iden-

tified potential confidentiality and integrity violations in TAP programs [3, 9, 21] and anti-patterns that could be leveraged in attacks to force devices into an insecure or unsafe state [6].

Research has also examined the prevalence of these potentially harmful anti-patterns. Some examined random or hand-crafted programs built from publicly available TAP rules [9, 21, 24], while others set up real devices to evaluate tools and test their attacks [6, 7, 15, 24]. Some prior work collected TAP programs from a small number of real users to evaluate [8, 11]. While many tools are capable of detecting various patterns, it is unclear whether users would understand what these patterns are, or if they find them undesirable. Therefore, we design tasks to measure understanding and perception of a selection of anti-patterns from prior work.

Usability of TAP tools The usability of a small number of proposed TAP analysis tools has been evaluated individually [15, 16, 28]; in contrast, we compare *techniques* employed by various tools from the TAP literature, examining the potential utility and efficacy of different approaches to specifying and debugging TAP programs and their properties.

3 Study Goals and Survey Design

We next describe the motivation for our study and how it informed the design of our surveys. We describe the specific procedures for administering the surveys in Section 4. Our goal is to examine the user-interaction approaches suggested by existing TAP analysis tools to determine what works well, where refinement is needed, and how tools might better serve users. To explore the breadth of the design space, we conduct a survey-based user study. We divided our study into four parts, which correspond to our research questions (Section 1).

3.1 Part 1: Identifying user needs

What are users trying to accomplish with their TAP programs? Since the space of tools and problems is large and varied, knowledge of users' goals can help inform priorities for tool design. For instance, if users are generally installing TAP rules without much consideration for what they expect their devices to do (or not do), they might find writing declarative policies difficult, and might instead benefit from a tool that looks for programming patterns that may cause problems.

We asked users about their high-level priorities for their TAP programs by rating the importance of several goals — Home Safety, Home Security, Privacy, Comfort and Convenience, Understanding Failures, and Fun — on a five-point Likert scale. We also asked them to describe any goals they had for their smart home in a free-response field.

3.2 Part 2: Writing declarative policies

Some tools allow users to specify how they want their devices to operate and check for violations of these declarative policies. There are a few ways to specify policies; an approach suggested by tools like AutoTap [28] and SafeTAP [16] is fill-in-the-blanks policy *templates*. However, these templates could be confusing to users, because they use complicated constructions such as "[state] should [always/never] be active while [state]" (see Table 1 for a full list).

So, we ask, can users correctly express their goals using this kind of interface? And, if not, what factors might be contributing to their problems? We break these questions down into the following three survey tasks.

Task 2a: Picking templates Can users pick a template format that matches a high-level goal? There is some variation between AutoTap [28] and SafeTAP [16] templates, but both tools have multiple templates that users need to pick from to specify their policies. While some templates are equivalent (like AutoTap's "_ should [always/never] be active" and SafeTAP's "I [always/never] want _"), there are important differences between others, like whether the policy is conditional or includes timing constraints, and other subtleties between (instantaneous) events and (persistent) states.

To investigate this, we presented participants with a scenario and goal, and asked them to choose which of the AutoTap/SafeTAP templates were most appropriate for the goal. Participants could choose any template from Table 1. Participants repeated this task for two scenarios.

Task 2b: Filling out templates Can users correctly fill out a policy template, once one has been correctly selected? To write policies understandable by a tool, a user may have to think of the goals they have for their home in terms of specific device behaviors or conditions since the tool likely will not understand what it means for a home to be "warm" or "safe." Some devices have simple boolean states, like "dryer on" or "dryer off," but others may require more specificity, like temperature or time.

Participants were presented a scenario describing a goal and a template, and their task was to fill in the blanks for template using drop-down menus of states or events. This task was repeated three times, for the following scenarios: closing the windows when it rains, setting the thermostat at night, and keeping the dryer off during work hours.

Task 2c: Clarity of templates Because there may be several equivalent ways to write a policy, we want to know which templates make most sense to users.

Policy templates themselves might be logically equivalent, like "_ should [always/never] be active" and "I [always/never] want _", or people might prefer one way of filling out a particular template over another, like "Dryer on should never be active" and "Dryer off should always be active." We are interested to know whether there are any templates that are especially popular, or if our participants tend to prefer equivalent AutoTap templates or SafeTAP templates, or the positive ("always") or negative ("never") form of the templates.

In this task, participants were shown a goal and four poli-

	Name	Template
AutoTap [28]	One-State Unconditional	[state] should [always/never] be active
	One-State Duration	[state] should [always/never] be active for more than [duration]
	Multi-State Unconditional	[state] and [state] should [always/never] occur together
	State-State Conditional	[state] should [always/never] be active while [state]
	Event-State Conditional	[event] should [only/never] happen when [state]
	Event-Event Conditional	[event] should [always/never] happen within [duration] after [event]
SafeTAP [16]	Whenever	Whenever [event] make sure that [state]
	Only When	[event] only when [state]
	Always/Never	I [always/never] want [state]

Table 1: Name and format of each policy template evaluated in the study.

cies describing the goal and were asked to pick the policy they felt was most natural. Participants were randomly shown three of five possible scenarios.

3.3 Part 3: Identifying TAP anti-patterns

Another class of TAP security analysis tools check for anti-patterns in TAP programs. An example of an anti-pattern is a loop, where one rule triggers a second rule, the second rule triggers the first, and so on.

We ask: Would users find these tools useful? How well can users identify anti-patterns on their own? Do they want assistance from a tool? And are there situations where users actually want to use these patterns, despite researchers having identified them as undesirable?

In this survey component, we investigate users' perceptions of anti-patterns identified by prior work, which we define in Table 2. We selected 12 anti-patterns from prior work and re-named them to reduce any bias the names and descriptions might introduce when evaluating their desirability (e.g., from "Action Conflict" [24] to "Opposite Behaviors"). We also added one that looks for any rules with different triggers and different actions (we call this "Different Triggers, Different Behaviors") to have something benign to compare against the anti-patterns identified by prior work as undesirable.

Participants were randomly assigned four anti-patterns. For each anti-pattern, we first provided participants with a definition and an example of the anti-pattern. Then, we asked participants to complete the following three tasks.

Task 3a: Identifying anti-patterns First, can users understand anti-patterns well enough to identify them in a TAP program, and do they have a good sense of which anti-patterns are more difficult to understand? In this task, we presented participants with four TAP programs, and asked them to select the program that was an instance of their assigned anti-pattern. We also asked them to rate the difficulty of understanding the anti-pattern on a five-point scale (Very Difficult, Difficult, Neither Difficult Nor Easy, Easy, Very Easy).

Task 3b: Perceptions of anti-patterns Do users believe anti-patterns to be problematic? Or, do they think there might be situations where people would want anti-patterns in their TAP programs? In this task, we asked participants four questions about their assigned anti-pattern: if they think their own TAP rules would contain the anti-pattern, if they would want the anti-pattern in their TAP rules, if they think others would want to use the anti-pattern, and if they want to avoid the anti-pattern. Participants responded on a four-point scale (Never, Rarely, Sometimes, Always).

Task 3c: Perceptions of tools for anti-patterns Would tools that detect anti-patterns be useful to users? We asked participants four questions about their assigned anti-pattern: whether they need help identifying it, whether they want help identifying it, whether they would use a tool that finds it, and whether they would be annoyed by a tool looking for the anti-pattern. Participants responded on a four-point scale (Definitely not, Probably not, Probably, Definitely).

3.4 Part 4: TAP program repair

TAP security analysis tools typically notify users of bugs or potential problems, but don't necessarily tell them how to fix them. Some tools describe the trace of events leading to the problem (e.g., Soteria [4]), others report only which rules are involved in the violation [16, 18], and yet others synthesize patches to suggest to the user [14, 28].

We ask, what types of information provided by analysis tools about an error or policy violation are most helpful for users trying to understand and correct the issue? In this survey component, we compare three forms of feedback.

Task 4a: Fixing a buggy rule Here we asked participants to fix a policy violation caused by a buggy rule. First, we showed participants a scenario and goal, and set of TAP rules, and asked them to pretend that a tool found a problem. In one scenario, the participants were told the user wants their door to be locked when they aren't home ("door lock" scenario), in another the user wants their lights to blink only to indicate that there is smoke in their home ("smoke", which is based on

Anti-Pattern Name	Description
Different Triggers, Same Behavior [4]	This pattern looks for rules that are triggered by different events, but lead to the same action.
Same Except No Condition [24]	This pattern looks for rules that are identical except that one has the WHILE condition and the other one doesn't.
Same Triggers, Different Behavior & Conditions [6, 7, 24, 27]	This pattern looks for two rules that are triggered by the same event and one rule has a WHILE condition and the other rule turns off the WHILE condition.
Chains with Opposite Behaviors [7, 24]	This pattern looks for rules that trigger other rules (i.e., form chains) and have different behaviors.
Chain [6, 7, 9, 19, 27]	This pattern looks for rules that may trigger other rules (i.e., form chains).
Different Triggers, Different Behaviors	This pattern looks for rules with different triggers and different behaviors.
Loops [1, 5, 7, 16, 24]	Triggering any rule in the loop will cause another rule to be triggered, which then causes the first rule to trigger again, leading to a loop.
Opposite Behaviors [1, 4, 5, 7, 15, 18, 19, 24, 27]	This pattern looks for rules that may trigger at the same time and cause opposite behaviors.
Same Behaviors [4, 5, 18, 24, 27]	This pattern looks for situations where multiple rules trigger the same behavior.
Un-Paired Rules [1, 12, 19]	Some rules form pairs (one rule might turn a device "on", while another turns it "off"). This pattern looks for rules that are missing their natural pair.
Extended Behavior [1, 12, 27]	This pattern looks for rules that do not account for behaviors that do not happen instantaneously, i.e., they are "extended" over a period of time.
Privacy [3, 5, 11, 18, 19, 21]	This pattern looks for rules that may allow people to learn private things about you.
Trust [5, 11, 18, 19, 21]	This pattern looks for rules that do things that require your trust.

Table 2: TAP anti-patterns included in S2, how we described them to participants, and the prior work inspiring them.

prior work [16]), and in the last scenario, the user never wants their house to be warmer than 72 degrees ("temperature").

Participants were randomly assigned to one of three conditions that determines how much additional information the "tool" gives them: a) which rule is involved, b) which rule is involved and the state of the home when the bug occurs, and c) a full trace that describes the series of events leading to the violation, including the initial state of all of the devices.

We then asked participants three questions to emulate the debugging process: (1) whether the problem is due to a missing rule, a (single) misbehaving rule, or interactions between multiple rules; (2) if the fix involves adding, modifying, or deleting a rule; and (3) to perform the fix, which involved writing a new TAP rule using an If-Then or If-While-Then template and drop-down menus, or choosing a rule to edit or delete (depending on their answer to the previous question). Participants repeated this task twice for two different scenarios, which were randomly assigned from three scenarios.

Task 4b: Fixing an incomplete program We test whether participants can fix a violation caused by a missing rule. Like the previous task, we showed participants a scenario, goal, and a set of TAP rules. Participants were randomly assigned to one of three conditions that determines how much additional information the "tool" gives them: a) a rule is missing; b) a rule is missing, and the state in which the missing rule is needed; and c) a full trace of events leading to the state in which the missing rule is needed. We then asked participants to write a TAP rule to fix the error, using an If-Then or If-While-Then template and drop-down menus. Everyone repeated this task for the same two scenarios.

4 Methodology

In this section, we describe the specific methods and procedures for conducting the study.

Survey structure Because of the length of the survey questions, we split our survey components across two different surveys to reduce the amount of time it took for individual participants to complete the study (see Figure 1 for an overview). Both surveys were implemented in Qualtrics.

Both surveys begin by asking participants to read and accept a consent form; then proceed to Part 1 (Identifying User Needs), which asks participants about their smart-home usage and their TAP goals. Then, to ensure that all participants have a baseline understanding of TAP, we give participants two practice exercises where we present them with a smart-home scenario, set a goal for them to achieve using home automations, and ask them to pick the appropriate TAP rule from a list of rules. They receive feedback explaining why their selection was or was not correct.

At this point, the survey flows diverge: Survey 1 (S1), proceeds to Tasks 2a-c (Writing Declarative Policies), while Survey 2 (S2) proceeds to Tasks 3a-c (Identifying TAP Anti-Patterns) followed by Tasks 4a-b (TAP Program Repair).

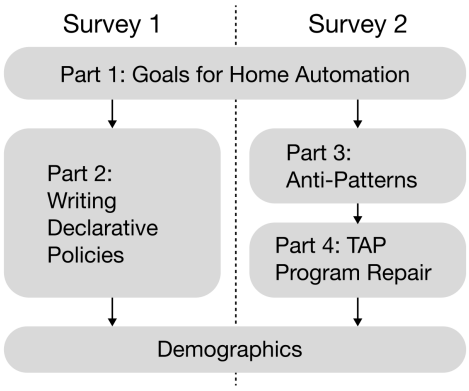


Figure 1: Diagram of the survey structure. Parts 2-4 were split across two surveys to shorten survey length.

At the end of both surveys, we collect participant demographics: age, gender, highest level of education achieved, and whether they have experience in a computing field. Lastly, we asked the participants for permission to publish their anonymized responses (available online [17]). We included two attention-check questions in each survey and discarded responses where both were answered incorrectly.

We revised the surveys over several pilot studies, which involved participants from varied technical backgrounds to help tune the difficulty of the tasks. The study was approved by the IRB at Carnegie Mellon University. Participants who completed a survey and passed at least one attention check received $10.00. The full text of both surveys is in Appendix B.

Recruitment We recruited participants with Prolific. Based on a rule-of-thumb sample-size estimation for an ordinal logistic regression we planned to conduct [23], we estimated we needed 175 participants for S1 and 275 for S2. Participants had to be 18 years or older, located in the US, fluent in English, and have experience with smart-home devices. Participants were only allowed to take one of the two surveys. We used Prolific's gender-balanced sample. S1 was published in October and S2 in November 2022. The median time to complete S1 was around 17 minutes, and 23 minutes for S2.

We received 176 complete responses for S1 and 278 responses for S2. We excluded responses that failed both attention check questions (1 for S1 and 3 for S2) as well as people who skipped more than 2 background questions about their smart-home experiences (1 for each survey). Responses were also evaluated for internal consistency (e.g., did any participants report no smart-home experience after passing the pre-screens?) and nonsensical text responses (1 for S2). In the end, we had 174 usable responses for S1 and 273 for S2.

Participant demographics Table 3 shows participant demographics for both surveys. Participants were balanced across gender, but skewed younger (only 16% and 18% were 46+ years old in S1 and S2, respectively). Most were educated, with 60% holding at least a 2-year degree across both surveys. Most participants did not have experience in a computing field (83%), and had used some form of automation in their own homes (61%).

	Survey 1		Survey 2	
	n	%	*n*	%
Gender				
Female	84	48.3%	131	48.0%
Male	87	50.0%	136	49.8%
Non-binary	3	1.7%	5	1.8%
Prefer to self-describe	0	0.0%	1	0.4%
Age				
18-25	39	22.4%	64	23.4%
26-35	63	36.2%	99	36.3%
36-45	44	25.3%	62	22.7%
46+	28	16.1%	48	17.6%
Highest Education Achieved				
Have not completed high school	0	0.0%	1	0.4%
High school or equivalent	68	39.1%	108	39.6%
Bachelor or associate degree	81	46.6%	130	47.6%
Graduate degree	24	13.8%	32	11.7%
Other	1	0.6%	2	0.7%
Experience in Computing?				
Yes	28	16.1%	47	17.2%
No	146	83.9%	226	82.8%
Experience with TAP?				
Yes	109	62.6%	164	60.1%
No	65	37.4%	109	39.9%
Total	174	100%	273	100%

Table 3: Demographics of study participants for both surveys.

We also collected data on the devices participants used in their homes. Almost all participants owned or used smart TVs (88%) and voice assistants (86%), but less than half owned non-entertainment-related devices like smart lights (50%) or thermostats (14%). The full list is available in Appendix C.

5 Results

This section presents our study results. The organization reflects the research questions and tasks from Sections 1 and 3.

5.1 RQ1: Users' Goals for TAP Rules

First, we present results on the goals that smart home users seek to accomplish with home automation.

Home safety and security are the most important high-level goals We asked participants to rate the importance of each of the following high-level goals for home automations: home safety, home security, comfort and convenience, understanding failures, privacy, and "just for fun", rating each on a five-point Likert scale. Figure 2 summarizes the responses.

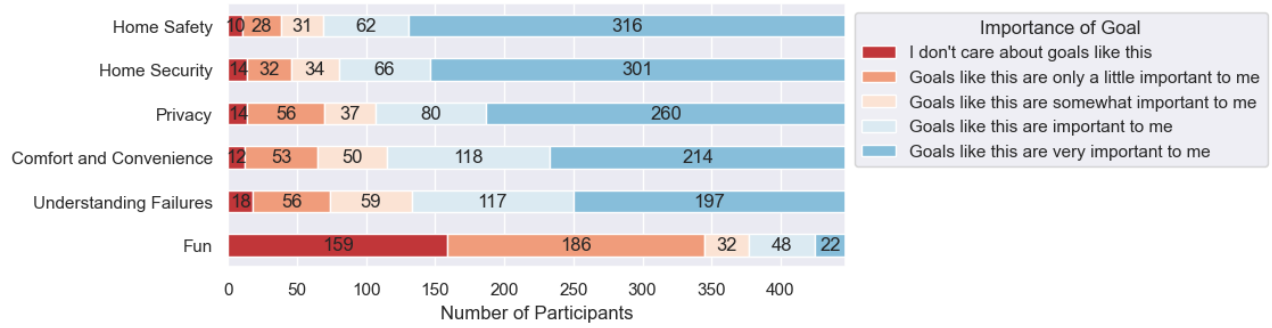


Figure 2: Importance of home automation goals. Safety and security were top goals, while fun was relatively unimportant. Each category included an example, like "Whenever my security camera turns off, I want to know why it happened" for Home Safety.

We found that the goals were divided into three tiers of importance: Home safety and security were the most important, with 84% and 82% of participants rating them as "important" or "very important". Privacy, comfort and convenience, and understanding failures were of secondary importance, with 76%, 74%, and 70% of participants rating them as (very) important. Notably, fun was not a strong motivation for using home automation, with 77% of participants reporting they don't care or that it is only a little important.

There were also 17 participants who described additional goals in a free response field. Some responses rephrased one of the above goals, but other goals included saving money/energy (5), accessibility (1), and health and safety (1).

Users' interest in TAP is primarily for comfort and convenience, home security We also asked participants in S1 to describe, in their own words, the purpose of the TAP programs in their home, or if they didn't have experience with TAP, what they hypothetically would do with TAP programs. We labeled their responses by the categories of goals identified above, and the level of specificity (at the whole home level vs. individual devices).

Contrary to the high-level self-reported goals, we found that the most common use-case for participants' TAP programs was comfort and convenience (mentioned in 69% of the responses) followed by home security (54%). Others mentioned saving money or energy (21%). Home safety, fun, privacy, and health appeared in fewer than 6% of the responses.

The difference between the importance of high-level goals and the actual programs that people write suggests that users want privacy and safety as an implicit property of their smart home, and will automate comfort/convenience or security related functionality using TAP.

Users goals for automation range in sophistication Responses ranged in specificity, indicating that users have different mental models for goals. Some responses (24%) were extremely broad, such as:

> *I want my home to be comfortable and secure. (P130)*

Many (59%) were specific to particular devices:

> *I would want my doors to be locked whenever I am not home. [...] certain devices to be turned on at certain times of the day with smart plugs, like my coffee machine for example. (P1)*

A few participants (16%) described both high-level goals and specific behaviors:

> *My goals are to keep my home safe and secure when I leave/while I am sleeping, so I want my doors locked and secure during these situations. I also love being able to control my thermostat from my phone. I also love using smart lights to help make my apartment look better and feel more like home. (P113)*

5.2 RQ2: Usability of Declarative Policies

Next, we explore the usability of template-based interfaces for specifying declarative policies about the desired state of the home, such as those proposed by AutoTap [28] and SafeTAP [16]. We evaluate whether users can a) pick a correct template format to achieve a goal and b) correctly fill out the fields of a template. We also investigate c) which templates seem most natural to users.

Task 2a: Participants can usually select a suitable template In this task, participants were presented two scenarios, and asked to pick a template that was appropriate for the scenario. In the *smoke detector* scenario, the home's lights should blink only when the smoke detector is triggered. In the second, *neighbors*, some automations cause the home's lights to blink, and the goal is that they should blink for at most 30 minutes to avoid annoying the neighbors. Participants chose from all nine SafeTAP and AutoTAP templates (Table 1).

In the *smoke detector* scenario, 91% of participants selected a valid template while 71% did likewise in the *neighbors* scenario. The correct templates for the first scenario included *Only When* (selected by 51%), *Event-State Conditional* (24%), *Whenever* (12%), or *Multi-State Unconditional*

(4%). Meanwhile, only *One-State Duration* (62%) or *Event-Event Conditional* (9%) could be used to write a policy for the *neighbors* scenario. The one-sample Pearson Chi-Squared tests indicated that the differences in the number of participants who picked each template were significantly different from chance (*smoke detector*: $\chi^2(8, N=174) = 290.37$, $p < .0001$, *neighbors*: $\chi^2(8, N=174) = 420.67$, $p < .0001$).

In general, it appears that participants can match a template to a high-level goal, but their success likely depends on the goal or situation: the more-complex *neighbors* scenario involved duration and had much lower success rates than the simpler *smoke detector* scenario.

Task 2b: Participants' success at filling out templates varies based on complexity of template and goal In this task, participants were assigned a scenario and a template, and were asked to *correctly fill-in-the-blanks* of the template to describe the goal. In the *window* scenario, the goal is to ensure the windows are closed when it rains; in *temperature*, that the room is cooler than 73F at night; and in *dryer*, that the dryer cannot run until after 5PM. In Table 4 we show participants' success at filling out their assigned templates.

We found that template filling success rates varied widely across scenarios and templates. For *window*, 74% of participants correctly filled out templates, compared to 28% for *temperature*, and 51% for *dryer*. Within each scenario, participants found more success with some templates than others. For example, 98% of participants assigned to the *Always/Never* template for *window* filled it out correctly, while only 37% were able to correctly complete the *Multi-State Unconditional* template for the same scenario. An analysis of variance based on mixed binomial logistic regression indicates a significant effect of scenario on correctness ($\chi^2(2, N = 522) = 47.9, p < 0.001$) and template on correctness ($\chi^2(6, N = 522) = 79.1, p < 0.001$).

We also observed that specific templates may be misinterpreted in certain situations. For *temperature*, no participants filled out the *Whenever* template correctly, which indicates that this template may be too confusing to use for duration conditions, even though it is technically adequate for the goal. There is also variation between the "always" and "never" forms of the same template. For instance, participants were more successful at filling out the *One-State Unconditional* template when they picked the "always" form (76% correct) than the "never" form (27%). In another case, some participants chose the "always" form of the *Multi-State Unconditional* template, even though it is not possible to correctly write the template in that form, meaning that 100% failed. This is consistent with prior work that observed that users tend to misinterpret the meaning of this template [28].

Task 2c: Template preferences vary Lastly, we investigated which templates participants preferred. We created five scenarios; in each we presented four sets of filled-in templates that satisfied a goal, and asked the participant to pick the one that sounded most natural to them. Within each scenario, we varied several factors, including the use of SafeTAP vs. AutoTap templates; always vs. never forms of templates; and whether multiple conditions were fulfilled via multiple templates or via one template using an AND or OR clause. The templates, their attributes, and the percentage of participants that chose each option are shown in Appendix E.

In each scenario, participants had somewhat clear preferences: the top two choices in each comprised over 75% of the votes. One-sample Pearson Chi-Squared tests confirmed that participants' choices were significantly different from chance.

The specific forms and attributes associated with the more popular templates varied from scenario to scenario. In 3 of 5 scenarios, SafeTAP templates were preferred over AutoTap. The *Whenever* template was relatively popular in each of the scenarios it appeared in (most popular in *security camera* and *forecast*, 2nd most popular in *smoke*), and the *Multi-State Unconditional* template was the least popular in both scenarios it appeared in. Participants did not prefer having fewer templates to more templates, nor was it clear if they preferred the "always" or "never" forms of templates.

5.3 RQ3: Understanding Anti-Patterns

Here, we report whether participants could identify TAP anti-patterns in S2 and whether they perceived them as undesirable. We see a large variation in participants' ability to identify anti-patterns and, surprisingly, that a substantial number of people may actually want to use anti-patterns in their programs.

Task 3a/b: Some anti-patterns are easy to spot, others are hard In this task, we showed participants the definition of an anti-pattern, and asked them to choose which of four example TAP programs contained that anti-pattern. Table 5 summarizes the results.

Overall, participants' success at identifying anti-patterns varies substantially between anti-patterns. Anti-patterns that most participants correctly identified typically involve redundant rules that share a trigger or action: *Same Behaviors* (75% correct), *Different Triggers, Same Behavior* (83%), and *Same Except No Condition* (83%). The anti-patterns participants had the most trouble identifying required understanding how long an action takes to complete (*Extended Behavior*, 34%), whether one rule can trigger another (Chains, 37%; and *Chains with Opposite Behaviors*, 31%), and other nuanced concepts, like the integrity of a trigger/action (*Trust*, 26%).

We also asked participants how difficult they found it to understand the anti-pattern they were tasked to look for. Figure 3 shows participants' responses. We again found a wide spread depending on the template. The easiest anti-pattern to understand was *Different Triggers, Same Behavior*, where 69% reported it was "easy" or "very easy" to understand, and the hardest was *Chains with Opposite Behaviors*, which 48% of people found "very difficult" or "difficult."

Participants' self-reported understanding of anti-patterns roughly correlates with their performance identifying the

Scenario	Template Name	Template	Overall % Correct	"Always" % Correct	"Never" % Correct
Window	Always/Never	I [always/never] want __	98	97	100
	One-State Unconditional	__ should [always/never] be active	91	97	71
	State-State Conditional	__ should [always/never] be active while __	67	64	86
	Multi-State Unconditional	__ and __ should [always/never] occur together	37	0	89
Temperature	One-State Unconditional	__ should [always/never] be active	62	76	27
	State-State Conditional	__ should [always/never] be active while __	39	41	29
	Multi-State Unconditional	__ and __ should [always/never] occur together	7	0	21
	Whenever	Whenever __ make sure that __	0	N/A	N/A
Dryer	State-State Conditional	__ should [always/never] be active while __	75	76	74
	Event-State Conditional	__ should [only/never] happen when __	63	0	96
	Whenever	Whenever __ make sure that __	57	N/A	N/A
	Event-Event Conditional	__ should [always/never] happen within __ after __	12	0	63

Table 4: Percent of participant responses that correctly filled out the blanks in a declarative policy template, across three scenarios. We also report the proportion of correct responses for the always/never form of each template, where applicable (the "only" form of event-state conditional is included under "always"). Success rates varied across scenarios and templates, indicating that people's ability to fill out policy templates is extremely context-specific.

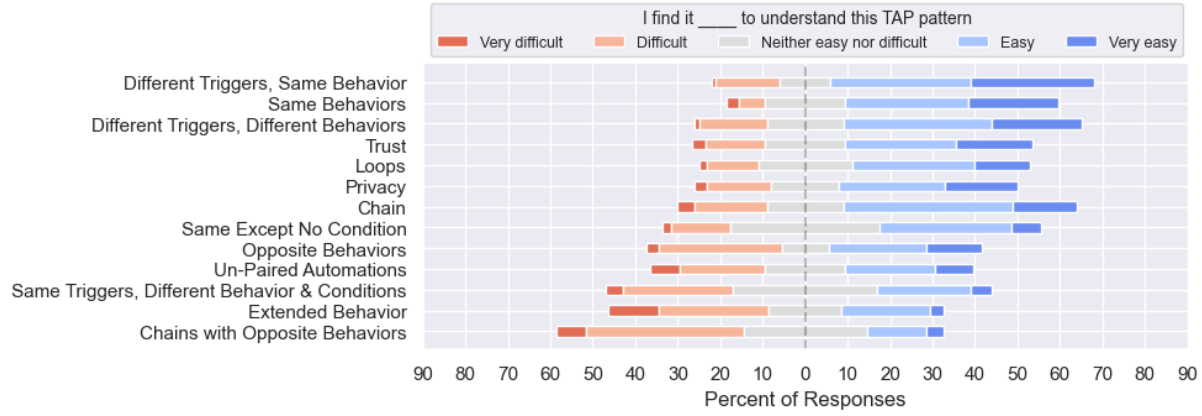


Figure 3: Perceived difficulty of identifying problematic TAP anti-patterns, sorted by average response score for each anti-pattern when converted to numerical values.

anti-pattern: an analysis of variance based on a mixed logistic regression indicated a significant effect of a participants' self-reported understanding of the anti-pattern and whether they correctly identified the template ($\chi^2(4,N=1683)=19.1$, $p<0.001$). However, for some specific anti-patterns, understanding and identification rates don't appear to align: for example, 58% of people thought that *Chain* was "easy" or "very easy" to understand, but only 37% of participants correctly identified the set of TAP rules exhibited that pattern.

Task 3b/c: Participants would accept a tool to find anti-patterns, but also *want* to use TAP anti-patterns We find that across anti-patterns, a majority of participants would "sometimes" or "always" like help identifying them (ranging from 57% to 78%) and would use a tool that helped them (66% to 89%). These responses (summarized in Appendix F) also roughly align with the perceived difficulty.

Surprisingly, participants did not find the anti-patterns universally undesirable. Figure 4 shows participants' perceptions of the desirability each anti-pattern. For all anti-patterns, many said they would "rarely" or "sometimes" *like to have* the anti-pattern in their home (at least 49% for each anti-pattern) and would want to *avoid* the anti-pattern "rarely" or "sometimes" (at least 51%). A majority also reported that their rules may "sometimes" or "rarely" have the anti-patterns (at least 67%).

However, a few anti-patterns were more undesirable than others: 41% of people always wanted to avoid *Opposite Behaviors*, 40% for *Chains with Opposite Behaviors* , and 31% for *Loops*. This suggests the TAP anti-patterns that researchers have identified as problematic might somehow be useful for people, or at least that people may not recognize them as problematic based on the description and example alone.

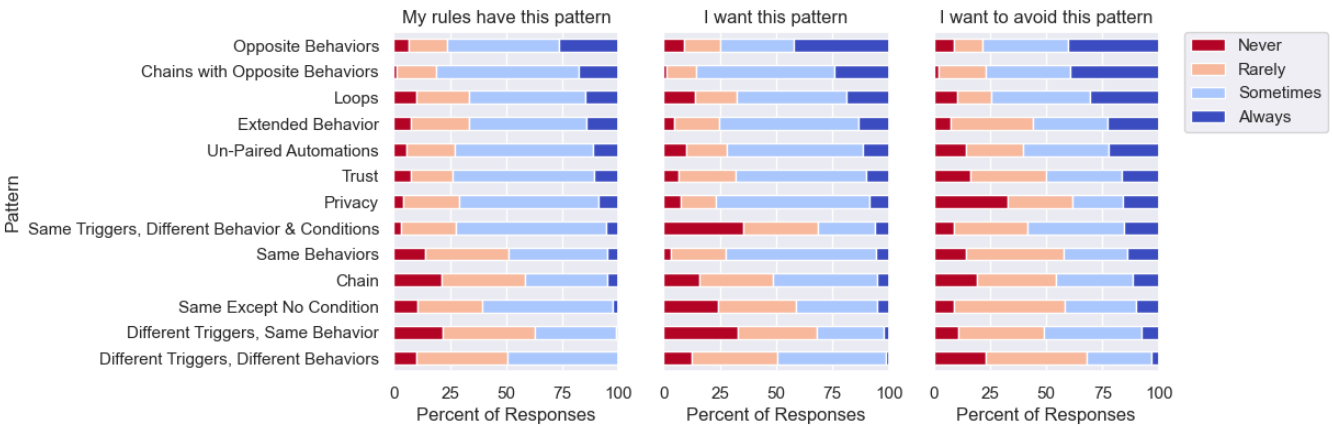


Figure 4: Participants' perceptions of the desirability of TAP patterns.

TAP Pattern Name	% Correctly Identified Round 1	Round 2
Loops	81%	47%
Same Behaviors	71%	80%
Privacy	67%	71%
Extended Behavior	52%	15%
Opposite Behaviors	51%	48%
Trust	48%	4%
Un-Paired Automations	30%	46%
Different Triggers, Same Behavior	83%	—
Same Except No Condition	83%	—
Different Triggers, Different Behaviors	53%	—
Same Triggers, Different Behavior & Conditions	57%	—
Chain	37%	—
Chains with Opposite Behaviors	31%	—

Table 5: Percent of participants that were able to identify the program that exhibited a TAP anti-pattern. For half of the anti-patterns, we tested participants a second time on an additional set of programs, shown in the rightmost column.

5.4 RQ4: Fixing TAP Programs

Many security tools, regardless of the types of properties they check, help users identify problems but do not fix them. Rather, tools typically generate some feedback, like a counterexample, to help the user in their efforts to make a fix. In this set of tasks, participants are asked to pretend a tool has identified some problem and they need to repair the program. In Task 4a, participants must identify and make modifications to fix a misbehaving rule, while in Task 4b, participants must add a missing rule. We vary the type of information the "tool" provides, showing either 1) which rule is misbehaving/that a rule is missing, 2) the rule and the state causing the issue, or 3) a full execution trace leading to the error, to investigate which form of feedback is mostly likely to help users.

Task 4a: Multiple types of feedback can help users repair broken rules We created three scenarios (*lock*, *temperature*, and *smoke*) where there is a violation of a declarative policy (worded like the goals in Section 5.2). In all of them, the error is caused by one incorrect rule and can be fixed by modifying or deleting it. We measure three stages of fixing the bug: identifying the problem, identifying how to fix it, and implementing the fix by writing a new rule or choosing the rule to modify/delete. Table 6 shows success rate of each fix stage for the feedback conditions across all three scenarios.

We found that the type of feedback had no effect on participants' ability to identify and fix bugs. For identifying the cause of the error, the difference in success rates ranged from 4-9% across conditions. The difference between conditions were also small for identifying how to fix (1-5%), and slightly larger for implementing a fix (9-15%). Analyses of variance based on mixed logistic regressions found no significant effect of condition on any of the three metrics ($\chi^2(2, N=546)=2.47$, *n.s.*, $\chi^2(2, N=546)=0.22$, *n.s.*, $\chi^2(2, N=546)=5.54$, *n.s.*).

However, the scenario did affect the success rate; namely, the *temperature* scenario was harder than the others. Across conditions, only 53% of participants successfully implemented a fix for *temperature*, versus 80% and 84% in for *lock* and *smoke*. The previous regressions indicated that scenario had a significant effect on success rates for all three metrics, ($\chi^2(2, N=546)=30.8$, $p<0.001$, $\chi^2(2, N=546)=41.2$, $p<0.001$, $\chi^2(2, N=546)=52.4$, $p<0.001$) and pairwise comparisons using Z-tests between scenarios in all three metrics showed *temperature* was significantly different from the others.

These results suggest that when a program is broken due to a single buggy rule, it is equally effective for tools to either indicate the rule causing the error (with or without additional information about the state of the home), or provide an execution trace of the error. They also suggest that some errors may be harder to fix than others, regardless of feedback provided.

Task 4b: State information helped participants write missing rules Unlike fixing bugs, we find that the type of

	Identified Problem			Identified Fix			Implemented Fix		
Condition	Lock	Smoke	Temp	Lock	Smoke	Temp	Lock	Smoke	Temp
Rule	84%	88%	71%	91%	95%	74%	84%	89%	63%
Rule w/State	86%	89%	68%	92%	93%	75%	75%	86%	48%
Full Trace	82%	88%	77%	92%	90%	77%	82%	77%	48%

Table 6: Percentage of participants that identified and implemented fixes successfully, across each scenario and tool condition. Different types of feedback did not have an effect on fix rates, but the temperature scenario was more difficult.

feedback does affect participants' success rate for adding missing rules. For both scenarios, participants were more likely to write the correct TAP rule when told that a rule is missing, and the state in which the program fails, (73% for *temperature*, 50% for *window*), rather only being told that a rule is missing (62% and 19%, respectively), or being given a full trace (57% and 20%).

An analysis of variance based on a mixed logistic regression found a significant effect of condition on success rate (χ^2(2,N=546)=23.5, p<0.001) and scenario on success rate (χ^2(2,N=546)=54.7, p<0.001). Post-hoc Z-tests indicate that the Missing Rule with State condition was significantly different from the other two conditions.

These results suggest that for the case where an error is caused by a missing rule, identifying the state which causes the issue makes it more obvious what rule to add, perhaps because it provides the user with a starting point.

5.5 Effect of TAP Experience

Surprisingly, we found that previous experience with TAP or computing, the number of smart home devices owned, and demographic factors had little correlation with participants' performance on the above tasks.

For Tasks 2b, 3a, 3b, 3c, 4a, and 4b, we conducted mixed logistic regressions to test if prior experience or demographics were correlated with the correctness of participants' answers or perceptions of the difficulty of a task. In each regression, we modeled the experimental conditions, participants' experience with TAP, experience with computing, number of devices owned, and demographic factors as predictors.

We only found a statistically significant effect for experience or demographics in three cases: In Task 4a, the success rate for implementing fixes to a template was 10 percentage points higher for participants with TAP experience (χ^2(1,N=546)=7.52,p=0.006), and the success rate for identifying fixes was 11 percentage points higher for male participants (χ^2(1,N=546)=5.61,p=0.017). In Task 2b, counterintuitively, the success rate for picking templates decreased by 4 percentage points for each smart home device a participant owned (χ^2(1,N=348)=4.61,p=0.032). The lack of consistent findings across tasks suggests that users that lack expertise in smart homes will not necessarily find security analysis tools harder to use, and that improving the usability of such tools is important even for expert users.

6 Discussion

Our study shows that trigger-action programming and security analysis tools for TAP have promise to be broadly accessible: participants worked with abstractions proposed in prior work successfully, e.g., to specify declarative policies for desired smart-home states and to identify anti-patterns in TAP programs. We also found that a high level of expertise in TAP or technology generally was not required to perform well at these tasks. At the same time, the results also suggest the need for more research: the features of some proposed tools were *much* less approachable than others (e.g., confusing declarative policy templates), and some of users' preferences clash with researchers' expectations (e.g., some users wanting to use anti-patterns). In this section, we make concrete recommendations based on our results, both for building tools and for future research.

6.1 Recommendations for Tool Developers

Templates are a promising approach for writing policies Encouragingly, participants were generally able to pick appropriate policy templates and fill them out, at least for some template formats in some scenarios (Section 5.2). This suggests that these template-based approaches to writing policies will work well for tools. The greatest confusion seems related to particularly confusing template formats or tricky scenarios, rather than to any fundamental misunderstanding of the interface, suggesting that with some refinement, these fill-in-the-blanks templates might be approachable to most users.

Policy tools should use context to guide users through template selection and filling Some of the more spectacular failures in the template-based tasks (Section 5.2) occurred when the template was not the most suitable choice for the context of the scenario, e.g., when non-duration based templates were assigned for duration-based scenarios. Though, more general-purpose templates technically can be used to satisfy duration-based goals, participants struggled with filling those templates. Thus, tools of this category could interac-

tively guide people into selecting more appropriate templates based on some contextual clues.

Analysis tools should provide specific information for debugging In many cases, participants were able to fix bugs in TAP programs when given sufficient information (Section 5.4). When a rule was broken, simply highlighting that specific rule was sufficient, but if a rule was missing, it was most helpful to also provide some context about when the rule would be needed. For security analysis tools, if automatically synthesizing a fix to a problem is not possible, providing information about the specific states and rules relating to the problem appears to be the best approach.

Tools should indicate how rules are modeled Some of the more confusing anti-patterns from Section 5.3, like *Extended Behavior*, require users to understand how TAP rules will behave in the real world. These results are consistent with prior work that identifies gaps in users' mental models of TAP programs and rules, especially when distinguishing instant (e.g., send email), extended (e.g., brew coffee), and sustained (e.g., turn lights on) behaviors [12]. In these cases, tools may need to provide more information about how the rule is modeled or interactivity to help people understand and fix the problem.

6.2 Recommendations for Researchers

Policy templates need refinement for comprehension In Section 5.2, we found several examples where specific templates had confusing wordings that negatively affected participants' ability to use them correctly. Among the most challenging to use templates are the "always" form of the "_ and _ should [always/never] occur together" template and the "Whenever _ make sure that _" template. The first was often used when it wasn't appropriate (which is consistent with prior findings [28]), and the confusion seemed to come from a mismatch between how the template is phrased (people appear to read it as "if-then") and the underlying formalism ("if-and-only-if"). Adjusting the wording of this template would likely improve its performance. On the other hand, the *Whenever* template is so popular (Table 13) and so frequently misused (Table 4), that it may be worth avoiding altogether.

Researchers developing tools to enforce declarative policy should carefully test the usability of the templates that they provide as an interface to users. It would also be helpful to dedicate some research to identify common characteristics of the most (and least) usable policy templates and develop guidelines for writing new policy templates. We can additionally leverage some of the insights from Section 5.1 to identify scenarios that research participants are more likely to find relevant and useful.

Users have an uneven grasp of anti-patterns, and tools could help TAP anti-patterns were tricky for people to understand and identify. Some were relatively simple and easy for people to understand, like *Different Triggers, Same Behavior*, which simply compare triggers an actions. In these cases, warnings with simple definitions and examples of the anti-pattern, like in our survey, could be sufficient.

Surprisingly, we also found that even anti-patterns which seem objectively undesirable (like *Loops* or *Conflicting Actions*, described in Section 1) were considered desirable (at least "sometimes") by many of our participants. This highlights the need for more research about how to communicate the threats posed by these anti-patterns, especially when users may overestimate their understanding of these anti-patterns.

6.3 Limitations

Our survey primarily used quantitative measures of usability, like task performance or rating scales for ease of use. We did not capture qualitative feedback on the usability of policy templates, anti-patterns, or template repair tools, which we leave to future work. We used vignettes to enable a controlled evaluation of templates and anti-patterns. However, the scenarios varied in difficulty and may not have been familiar to all participants, which could have contributed to participants' poor performance on some tasks. Because these tasks were presented as vignettes in a survey interface, the findings may not generalize to a live smart-home setting, due to differences in the user interfaces and the scenarios in which users would encounter anti-patterns or create templates in practice. Not all participants in our study had prior experience with TAP (39% did not), and for those that did, we did not characterize their level of expertise with TAP. Though we did not find an effect of prior experience on most tasks, in real deployments familiarity with smart-home configuration and TAP may impact the usability of security analysis tools.

7 Conclusion

In this paper, we presented the results of our exploratory survey of TAP security analysis approaches. We found considerable variation in the success and perceived utility of various approaches. Participants were generally capable of picking the correct templates for implementing specified high-level policies; however, while they generally filled out some templates very accurately, there were other templates where almost all participants struggled. We found that participants were more successful at debugging TAP programs when they knew some of the relevant state conditions involved in the violation, compared to only telling them which rule was involved or sharing all of the events leading up to the violation. Participants had more difficulty identifying some anti-patterns than others, didn't always seem to realize when they were having difficulty understanding them, and didn't find any of them wholly undesirable. More research is needed to determine how to best describe anti-patterns to facilitate understanding and communicate the threats they pose.

Acknowledgments We would like to express our gratitude to the numerous students, staff, faculty, and friends who helped pilot our studies and offered helpful feedback to improve their design. This work was supported in part by Carnegie Mellon CyLab, Cisco Research through the Carnegie Mellon CyLab partnership program, NSF award CNS2114148, and a CyLab Presidential Fellowship.

References

[1] Will Brackenbury, Abhimanyu Deora, Jillian Ritchey, Jason Vallee, Weijia He, Guan Wang, Michael L. Littman, and Blase Ur. How users interpret bugs in trigger-action programming. In *Proc. of CHI*, 2019.

[2] Lei Bu, Wen Xiong, Chieh-Jan Mike Liang, Shi Han, Dongmei Zhang, Shan Lin, and Xuandong Li. Systematically ensuring the confidence of real-time home automation IoT systems. *ACM Trans. on Cyber-Physical Systems*, 2, 2018.

[3] Z. Berkay Celik, Leonardo Babun, Amit K. Sikder, Hidayet Aksu, Gang Tan, Patrick McDaniel, and A. Selcuk Uluagac. Sensitive information tracking in commodity IoT. In *Proc. of USENIX Security*, 2018.

[4] Z. Berkay Celik, Patrick McDaniel, and Gang Tan. SOTERIA: Automated IoT safety and security analysis. In *Proc. of USENIX ATC*, 2018.

[5] Z Berkay Celik, Gang Tan, and Patrick D McDaniel. IoTGuard: Dynamic enforcement of security and safety policy in commodity IoT. In *Proc. of NDSS*, 2019.

[6] Haotian Chi, Chenglong Fu, Qiang Zeng, and Xiaojiang Du. Delay wreaks havoc on your smart home: Delay-based automation interference attacks. In *Proc. of IEEE SP*, 2022.

[7] Haotian Chi, Qiang Zeng, Xiaojiang Du, and Jiaping Yu. Cross-app interference threats in smart homes: Categorization, detection and handling. In *Proc. of IEEE/IFIP DSN*, 2020.

[8] Camille Cobb, Milijana Surbatovich, Anna Kawakami, Mahmood Sharif, Lujo Bauer, Anupam Das, and Limin Jia. How risky are real users' IFTTT applets? In *Proc. of USENIX SOUPS*, 2020.

[9] Wenbo Ding and Hongxin Hu. On the safety of IoT device physical interaction control. In *Proc. of ACM CCS*, 2018.

[10] Home Assistant. Home Assistant: Awaken your home. https://www.home-assistant.io, 2023.

[11] Kai-Hsiang Hsu, Yu-Hsi Chiang, and Hsu-Chun Hsiao. SafeChain: Securing trigger-action programming from attack chains. *IEEE Transactions on Information Forensics and Security*, 14(10), 2019.

[12] Justin Huang and Maya Cakmak. Supporting mental model accuracy in trigger-action programming. In *Proc. of ACM UbiComp*, 2015.

[13] IFTTT. IFTTT: Every thing works better together. https://ifttt.com, 2023.

[14] Chieh-Jan Mike Liang, Lei Bu, Zhao Li, Junbei Zhang, Shi Han, Börje F. Karlsson, Dongmei Zhang, and Feng Zhao. Systematically debugging IoT control system correctness for building automation. In *Proc. of ACM BuildSys*, 2016.

[15] Chieh-Jan Mike Liang, Börje F. Karlsson, Nicholas D. Lane, Feng Zhao, Junbei Zhang, Zheyi Pan, Zhao Li, and Yong Yu. SIFT: Building an internet of safe things. In *Proc. of IPSN*, 2015.

[16] McKenna McCall, Faysal Hossain Shezan, Abhishek Bichhawat, Camille Cobb, Limin Jia, Yuan Tian, Cooper Grace, and Mitchell Yang. SafeTAP: An efficient incremental analyzer for trigger-action programs, 2021. CMU Technical Report.

[17] McKenna McCall, Eric Zeng, Faysal Hossain Shezan, Mitchell Yang, Lujo Bauer, Abhishek Bichhawat, Camille Cobb, Limin Jia, and Yuan Tian. Supplementary material. https://kilthub.cmu.edu/articles/dataset/Towards_Usable_Security_Analysis_Tools_for_Trigger-Action_Programming_--_Dataset/23100482, 2023.

[18] Dang Tu Nguyen, Chengyu Song, Zhiyun Qian, Srikanth V. Krishnamurthy, Edward J. M. Colbert, and Patrick McDaniel. IotSan: Fortifying the Safety of IoT Systems. In *Proc. of CoNEXT*, 2018.

[19] Mitali Palekar, Earlence Fernandes, and Franziska Roesner. Analysis of the susceptibility of smart home programming interfaces to end user error. In *Proc. of IEEE SPW*, 2019.

[20] SmartThings Inc. SmartThings: One simple home system. A world of possibilities. https://www.smartthings.com, 2023.

[21] Milijana Surbatovich, Jassim Aljuraidan, Lujo Bauer, Anupam Das, and Limin Jia. Some Recipes Can Do More Than Spoil Your Appetite: Analyzing the Security and Privacy Risks of IFTTT Recipes. In *Proc. of WWW*, 2017.

[22] Blase Ur, Elyse McManus, Melwyn Pak Yong Ho, and Michael L. Littman. Practical trigger-action programming in the smart home. In *Proc. of SIGCHI*, 2014.

[23] Maarten van Smeden, Karel GM Moons, Joris AH de Groot, Gary S Collins, Douglas G Altman, Marinus JC Eijkemans, and Johannes B Reitsma. Sample size for binary logistic prediction models: Beyond events per variable criteria. *Statistical Methods in Medical Research*, 28(8), 2019.

[24] Qi Wang, Pubali Datta, Wei Yang, Si Liu, Adam Bates, and Carl A. Gunter. Charting the attack surface of trigger-action IoT platforms. In *Proc. of ACM CCS*, 2019.

[25] Qi Wang, Wajih Ul Hassan, Adam Bates, and Carl Gunter. Fear and logging in the internet of things. In *Proc. of NDSS*, 2018.

[26] Svetlana Yarosh and Pamela Zave. Locked or not? mental models of IoT feature interaction. In *Proc. of CHI*, 2017.

[27] Yinbo Yu and Jiajia Liu. TAPInspector: Safety and liveness verification of concurrent trigger-action IoT systems. *IEEE Trans. on Information Forensics and Security*, 17, 2022.

[28] Lefan Zhang, Weijia He, Jesse Martinez, Noah Brackenbury, Shan Lu, and Blase Ur. AutoTap: Synthesizing and repairing trigger-action programs using LTL properties. In *Proc. of IEEE/ACM ICSE*, 2019.

A Terminology

The terminology differs slightly between our paper and the surveys. We summarize the relevant terminology in Table 7.

B Survey Instrument

This section includes the survey instrument for both S1 and S2. Both surveys include the following sections:

- Consent form, pre-screen questions, and Prolific ID collection
- Background questions on smart home devices; warm-up exercises introducing TAP rules; home automation goals
- Main survey tasks
- Demographic and other wrap-up questions

The surveys are the same except where indicated. We include question text (shortened for brevity) and describe the survey flow decisions and other details about the survey using *italics*, fields from tables using **bold text**, and list answer choices next to circles: ○

Background and warm-up (RQ1) *Introduction to the background questions.*

1. For each of the following categories, tell us if you have used and/or have heard of the device. If you own or have set up a device not listed here, you can use the box at the bottom to describe the device. *Device categories are shown in Figure 12*
   ○ I have used smart features on one of these ○ I have used one of these, but not the smart features ○ I have heard of these, but never used one ○ I have never heard of these

2. *For the devices the participant reports they have used/heard of in Question 1:* For each of the following categories, tell us if you own and/or have set up the device. If you own or have set up a device not listed here, you can use the box at the bottom to describe the device.
   ○ I own and have set up one of ○ I own one of these, but have not set it up ○ I do not own one of these but I have set one up ○ I do not own, nor have I set up one of these

Practice exercises to introduce home automations

3. Have you ever tried to use home automations with your own smart home devices? (Please select "yes" even if you tried to set up home automations without success.) *Allows multiple answers*
   ○ Yes, using a platform like IFTTT, SmartThings, HomeKit, or openHAB ○ Yes, using built-in settings like schedules or based on my location ○ No

4. Which of the following reasons stop you from using more smart home devices than you currently do? (either installing more home automations with your current devices, buying more devices, or using more features on your current devices).
   ○ This is not a factor ○ This is somewhat a factor ○ This is definitely a factor

   (a) Learning to use a new device/feature/home automation is too difficult
   (b) Cost is too high
   (c) Not interesting to me
   (d) Privacy and/or security concerns
   (e) Other (please describe below) *Free response*

5. *S1 only. Wording depends on whether the participant indicated they have used home automations before in Question 3.* Which of the following describes your goals for your home automation? If you have different goals for different types of home automation, you can select multiple options. *Allows multiple answers*

Term	Survey term	Definition
TAP rule	Home automation	Trigger-action programming (TAP) rules allow users to customize their smart home devices. A simple TAP rule format is "IF *trigger* THEN *action*" which causes *action* to happen in response to the *trigger*. In the surveys, we often use the word *behavior* instead of *action*.
TAP program	Multiple home automations	A set of TAP rules which may (or may not) interact with each other when one rule triggers another rule.
(Declarative) policy template	Template	Tools which allow users to check that their smart homes perform specific behaviors use policy templates to help the user communicate their goals to the tool.
TAP patterns	Patterns	Tools which look for trigger-action programming patterns are checking for potentially dangerous/undesirable interactions between TAP rules rather than asking users to specify their own custom property.

Table 7: Terms used in this paper and their survey counterparts.

○ I have goals which some of my home automations work together to achieve ○ I have independent goals which my home automations would achieve independently ○ I have no specific goals in mind for some of my home automations

6. *S1 only.* Can you describe the goals you had in mind (if any) when you were answering the last question? *Free response*

7. Please select from the choices below to tell us how important these goal categories are to you.
   ○ Goals like this are very important to me ○ Goals like this are somewhat important to me ○ Goals like this are only a little important to me ○ I don't care about goals like this
   Categories are listed in Figure 2

8. Did you think of any other goal(s) which did not fit into the categories we identified, but would be important to you?
   ○ Yes ○ No

9. *If the participant answered "Yes" in Question 8.* How would you describe the goal(s)? *Free response*

S1 tasks (RQ2)

*We introduce the concept of **declarative policies** and the **templates** used to write them. Policy templates and the work they come from are shown in Table 1. We write templates with **bold text** and underline fields in the templates. Fields which have not been filled are written __. For each task we specify whether the participant chooses a template (which may be filled or unfilled), or fill out the template using a drill-down interface.*

Participants are randomly assigned a survey flow to account for the learning effect of S1 Task 1: half are shown Task 1 first and the rest are shown Task 1 last.

S1 Task 1: Template preference In this task, we want to know which goal formats are the most natural to use.

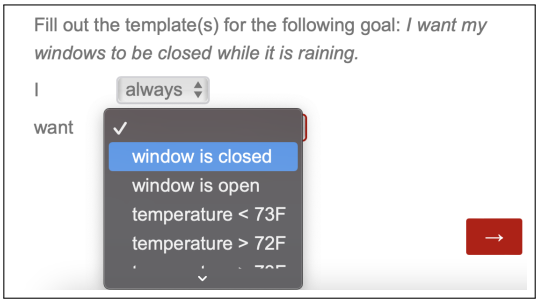


Figure 5: Screenshot of the interface for Task 2a, where the participants are asked to fill out the "**I** always/never **want** __" policy template for the "windows" scenario.

In the following questions, please select the option which best describes the goal, in your opinion. If none of the choices seem natural, please select the last option and briefly explain why. *Participants are shown 3 of the 5 questions in this section. Full question text may be found in our supplementary material [17].*

S1 Task 2: Template picking *Participants are shown both of the following questions and pick from one of the unfilled templates shown in Figure 1*

15. Pick one template for the following goal: I want my lights to blink to tell me that smoke has been detected. If the lights are blinking, I will assume there is a fire, so I don't want them to blink for any other reason.

16. Pick one template for the following goal: I have some automations to cause my lights to blink, but I am worried they might blink all night long and disturb my neighbors while I am out of town. I want to know that when they blink, they never blink for more than half an hour.

S1 Task 3: Template filling *Participants are shown each of the following questions and one of the templates (randomly selected), which they fill via a drill-down interface (shown in*

Figure 5). The participant also picks between the underlined choices shown in brackets.

17. Fill out the template(s) for the following goal: I want my windows to be closed while it is raining.
    (a) **I** [always/never] **want __**
    (b) **__ and __ should** [always/never] **occur together**
    (c) **__ should** [always/never] **be active**
    (d) **__ should** [always/never] **be active while __**
18. Fill out the template(s) for the following goal: I am worried my baby will overheat at night. I read that they should sleep in rooms cooler than 73F.
    (a) **Whenever __ make sure that __**
    (b) **__ and __ should** [always/never] **occur together**
    (c) **__ should** [always/never] **be active**
    (d) **__ should** [always/never] **be active while __**
19. My dryer is disrupting my Zoom meetings. I don't want it to run during business hours. My meetings are always finished by 5PM.
    (a) **Whenever __ make sure that __**
    (b) **__ should** [always/never] **be active while __**
    (c) **__ should** [only/never] **happen when __**
    (d) **__ should** [always/never] **happen within __** [minutes/hours] **after __**

S2 tasks (RQ3 & RQ4)

Introduction to general properties, which we refer to as "patterns" in the survey.

TAP Pattern task 1 *Each participant answers the following questions for 2 properties, randomly selected from Table 8. The section begins with an explanation of **[Pattern]**, including an example. Half of the participants are asked to identify the example of **[Pattern]** first, the other half are asked the Likert questions first.*

10. Please select the example of automations with the "**[Pattern]**" pattern. (Exactly 1 will fit the pattern.)
    Participants choose from 4 simple programs, the correct response is shown in Table 8.
11. How difficult do you think it is to understand what the "**[Pattern]**" pattern is, as a concept?
    ○ Very easy ○ Easy ○ Neither easy nor difficult ○ Difficult ○ Very difficult
    (a) I find understanding **[Pattern]** to be...
    (b) Technical people would find understanding **[Pattern]** to be...
    (c) Most people would find understanding **[Pattern]** to be...
    (d) Finding **[Pattern]** in my home automations would be...
12. How often would people have the "**[Pattern]**" pattern in their home automations? How often would people want to have the "**[Pattern]**" pattern in their home automations?
    ○ Always ○ Sometimes ○ Rarely ○ Never
    (a) I think my home automations would have **[Pattern]**...
    (b) I would want **[Pattern]** in my home automations...
    (c) I would want to avoid **[Pattern]** in my home automations...
    (d) I think other people would want **[Pattern]** in their home automations...
13. Would people want help from a tool to find the "**[Pattern]**" pattern in their home automations?
    ○ Definitely ○ Probably ○ Probably Not ○ Definitely Not
    (a) I would need help finding **[Pattern]** in my home automations...
    (b) I would want help finding **[Pattern]** in my home automations...
    (c) I would use a tool if it looked for **[Pattern]**...
    (d) It would be annoying if the tool looked for **[Pattern]**...

TAP Pattern task 2 *This task is similar to the previous one except that the participant is asked to identify 2 examples of each **[Pattern]**. Each participant is shown 2 properties, randomly selected from Table 9. The section begins with an explanation of **[Pattern]**, including an example. Half of the participants are asked to identify the example of **[Pattern]** first, the other half are shown the Likert questions first.*

Bug fixing task

For this task, participants are randomly assigned to one of three groups which determines how much feedback they get to help them repair the bugs in this task: rule only, rule and state, or a full trace evets leading to the bug. Each participant is given 2 programs to fix. The scenario, program, and feedback given to each group is shown in Table 10.

Next, we are going to ask questions about what actions you would take if a tool reported possible problems with your home automations. Suppose you told a tool that your goal is **[Scenario]**. You have the following home automations installed: **[Buggy program]**.

14. What problem is happening here?
    ○ A home automation is missing ○ A home automation is

Property	Program
Different Triggers, Same Behavior [4]	“IF user arrives at home THEN lock door” “IF user leaves home THEN lock door”
Same Except No Condition [24]	"IF it begins raining THEN close window" "IF window is opened WHILE it is raining THEN close window"
Same Triggers, Different Behavior & Conditions [6, 7, 24, 27]	“IF the temperature is >74F WHILE the A/C is off THEN turn on A/C” “IF the temperature is >72F THEN turn on A/C”
Chains with Opposite Behaviors [7, 24]	“IF temperature <74F THEN turn off A/C” “IF temperature >78F THEN turn on A/C”
Chain [6, 7, 9, 19, 27]	"IF temperature >78F THEN turn on A/C" "IF temperature <73F THEN turn off A/C"
Different Triggers, Different Behaviors	"IF it begins raining THEN set light color to blue" "IF it becomes sunny THEN set light color to yellow"

Table 8: Example of each pattern for S2. We include citations for patterns which are inspired by prior work.

Anti-Pattern Name		Program
Loops [1, 5, 7, 16, 24]	(1)	"IF I post a new Facebook status THEN post a new tweet" "IF I post a new tweet THEN post a new Facebook status"
	(2)	"IF temperature >70F WHILE A/C is off THEN turn on A/C" "IF temperature <73F WHILE A/C is on THEN turn off A/C"
Opposite Behaviors [1, 4, 5, 7, 15, 18, 19, 24, 27]	(1)	"IF motion is detected THEN turn on security camera" "IF the time is 6AM THEN turn off security camera"
	(2)	"IF temperature >70 THEN turn on A/C" "IF temperature <73 THEN turn off A/C"
Same Behaviors [4, 5, 18, 24, 27]	(1)	"IF temperature >74F THEN turn on A/C" "IF humidity >80% THEN turn on A/C"
	(2)	"IF it begins raining THEN close window" "IF window is opened WHILE it is raining THEN close window"
Un-Paired Automations [1, 12, 19]	(1)	"IF it begins raining THEN close window" "IF window is opened WHILE it is raining THEN close window"
	(2)	"IF a presence is detected THEN turn on security camera" "IF new reminder added THEN send notification"
Extended Behavior [1, 12, 27]	(1)	“IF the time is 7AM THEN begin brewing coffee” “IF user arrives at home THEN begin brewing coffee”
	(2)	"IF email is received THEN turn on light sequence" "IF user leaves work THEN turn off light"
Privacy [3, 5, 11, 18, 19, 21]	(1)	“IF user leaves home THEN turn on porch light” “IF user arrives at home THEN unlock door”
	(2)	"IF the time is 7AM THEN post a new tweet" "IF I post a new status on Facebook THEN post a new tweet"
Trust [5, 11, 18, 19, 21]	(1)	"IF someone I follow tags me THEN re-tweet their post” "IF I post a new status on Facebook THEN post a new tweet"
	(2)	"IF an email is received THEN flash lights" "IF new reminder added THEN send notification"

Table 9: Example of each pattern for S2. Participants are asked to identify the example of **[Pattern]** twice for this task. We include citations for patterns which are inspired by prior work.

	Scenario text	Buggy program
S1	Whenever I am not at home, I want my door to be locked	**"IF the time is 7AM THEN unlock door"** "IF the user leaves home THEN lock door" "IF the temperature is above 75 THEN turn on A/C" "IF the temperature is below 68F THEN turn off the A/C"
S2	I want my lights to blink only when smoke has been detected	**"IF new email received THEN blink lights"** "IF smoke detected THEN blink lights" "IF the time is 7AM THEN unlock door" "IF the user leaves home THEN lock door"
S3	Temperature above 72 should never happen	**"IF window opened THEN turn off the A/C"** "IF the time is 7AM THEN unlock door" "IF the temperature is below 68F THEN turn off the A/C" "IF the temperature is 71F THEN turn on the A/C"

	State	Full trace
S1	the time is 7AM	Initially, the time is 6AM, the user is at home, the door is unlocked, the temperature is 70F, and the A/C is off. Next, the time is 6:30AM and the user leaves home. An automation is triggered and the door is locked. Finally, the time is 7AM. An automation is triggered and the door is unlocked.
S2	new email received	Initially, the time is 6AM, the user is at home, the door is unlocked, the lights are not blinking, and no smoke is detected. Next, the time is 6:30AM and an email is received. An automation is triggered and the lights begin blinking.
S3	window opened	Initially, the time is 10AM, the window is closed, the door is unlocked, the temperature is 70F, and the A/C is off. The temperature is rising. Next, the time is 10:30AM, the temperature is 71F, and the user opens the window. An automation is triggered and the A/C turns on. The temperature begins falling. An automation is triggered and the A/C is turned off. The temperature begins rising. The time is 11AM and the temperature is 72F. Finally, the time is 11:30AM and the temperature is 73F.

Table 10: Bug-fixing task for S2. For each scenario we show the full program, and the feedback given to the participant to help them make a fix. The rule shown in bold text is the one shown to participants as feedback.

	Scenario text	Partial program
S1	The window should never be open while it is raining	"IF the user leaves home THEN lock door" "IF it begins raining THEN close window" "IF temp > 75 WHILE it is not raining THEN open window" "IF temp > 75 WHILE it is raining THEN turn on A/C"
S2	The temperature should never be above 75F for more than 1 hour	"IF it begins raining THEN close window" "IF the user leaves home THEN lock door" "IF smoke detected THEN blink lights" "IF the time is 7AM THEN unlock door"

	State	Full trace
S1	user opens window and it is raining	Initially, the temperature is 70F, the window is open, and it is not raining. Next, it begins raining. An automation is triggered, closing the window. Finally, the user opens the window.
S2	the temperature increases above 75F	Initially, the time is 10AM, the temperature is 75F, the window is open, and it is not raining. The temperature is rising. Next, the time is 10:30AM and the temperature is 76F. Next, the time is 11AM and the temperature is 77F. Finally, the time is 11:30AM and the temperature is 78F.

Table 11: Program writing task for S2. For each scenario we show the partial program, and the feedback given to the participant to help them complete the program.

doing something I don't want ○Multiple home automations are interacting to do something I don't want ○I don't know ○Something else *Free response*

15. What action would you take?
○Add another home automation ○Modify a home automation ○Delete a home automation ○Do nothing/I don't know ○Something else *Free response*

16. *If "Add another home automation" was selected for Question 15* Which format would the new home automation take?
○If-then ○If-while-then

17. *Wording depends on whether "If-then" or "If-while-then" was selected for Question 16* Please enter the missing fields for the "if-then" home automation, below. *Drill-down format*

18. *Wording depends on whether "Modify a home automation" or "Delete a home automation" was selected for Question 15* Which home automation would you modify? *Participant selects one of the TAP rules in* ***[Buggy program]***

TAP program completion task *For this task, participants are assigned to a group which determines how much feedback they get to help them complete the partial TAP program. They are assigned to the same group as the previous task. Each participant is given the same 2 programs to complete. The scenario, program, and feedback given to each group is shown in Table 11.*

Suppose you told a tool that your goal is **[Scenario]**. You have the following home automations installed: **[Partial program]**.

If the participant is given a rule as feedback: If the tool told you there was a missing home automation, what home automation would you add?

If the participant is given a rule and state as feedback: If the tool told you there was a missing home automation when: **[State]** What home automation would you add?

If the participant is given a full trace as feedback: If the tool told you there was a missing home automation after the following sequence of events: **[Full trace]** What home automation would you add?

19. Would you like to add a home automation in the "if-then" or the "If-while-then" format?
○If-then ○If-while-then

20. *Wording depends on whether "If-then" or "if-while-then" was selected for Question 19* Please enter the missing fields for the "if-then" home automation, below. *Drill-down format*

Smart Devices Owned	S1	S2
Voice Assistant	81.6%	78.8%
Smart TV	78.7%	83.5%
Smart Lightbulb or Switch	36.2%	44.7%
Doorbell Camera	28.2%	35.9%
Security Camera	27.6%	39.2%
Smart Thermostat or A/C	26.4%	23.8%
Smart Vacuum or Mop	12.6%	17.9%
Smart Lock	9.8%	8.8%
Baby Monitor	9.2%	13.2%
Other	7.5%	8.8%
Smart Smoke or CO Detector	5.7%	5.1%
Smart Lawn Mower or Sprinkler	0.6%	2.9%

Table 12: Percent of participants that own smart home devices from various categories.

C Device Background and Experience

Table 12 shows the percent of participants that own each type of smart home device. Most participants have a smart TV or smart speaker. Other device types are less common.

D Statistical Results for TAP Goals (RQ1)

To determine which goals for TAP were most important to participants, we conducted 15 pairwise Wilcoxon signed-rank tests, corrected with Holm's sequential Bonferroni procedure, between each pair of goals. In each test, we paired each participants' responses to one goal to their response in the other goal. The tests indicated significant differences in the proportion of responses for every pair of goals except for (Home Security, Home Safety), (Comfort and Convenience, Understanding Failures), and (Comfort and Convenience, Privacy).

E Template Preferences (RQ2)

Table 13 shows the proportion of responses for each of the policies in the template preferences task as well as the templates involved in each policy and their attributes.

F Anti-pattern preferences (RQ3)

Table 14 shows the proportion of participants who report needing/wanting help identifying TAP anti-patterns, as well as how many would use a tool that looks for the anti-patterns or would be annoyed by a tool looking for the anti-patterns.

Scenario	Tool	Template(s)	Policy Structure	Sentiment	% Preferred
Temperature	AutoTap	State-State Conditional	2 Templates	Negative	42%
	SafeTAP	Always/Never	1 Template With And	Positive	38%
	SafeTAP	Always/Never	1 Template With Negation + Or	Negative	13%
	N/A	None of the Above	N/A	N/A	5%
	AutoTap	One-State Duration	2 Templates	Negative	2%
Smoke	AutoTAP	Event-State Conditional	1 Template	Positive	50%
	SafeTAP	Whenever	2 Templates	Positive	24%
	SafeTAP	Only When	1 Template	Positive	21%
	AutoTap	Multi-State Unconditional	2 Templates	Negative	4%
	N/A	None of the Above	N/A	N/A	1%
Security Camera	SafeTAP	Whenever	2 Templates	Positive	67%
	Both	Whenever, Event-State Conditional	3 Templates	Negative	16%
	AutoTap	State-State Conditional	1 Template With Or	Positive	11%
	AutoTap	Multi-State Unconditional	2 Templates	Negative	5%
	N/A	None of the Above	N/A	N/A	1%
Humidity	SafeTAP	Always/Never	1 Template With And	Positive	46%
	SafeTAP	Always/Never	2 Templates	Negative	29%
	AutoTap	One-State Unconditional	1 Template With And	Positive	14%
	AutoTap	One-State Unconditional	2 Templates	Negative	9%
	N/A	None of the Above	N/A	N/A	2%
Forecast	SafeTAP	Whenever	Consistent Conjunctions	Positive	42%
	SafeTAP	Only When, Whenever	Mixed Conjunctions	Positive	25%
	Both	Whenever, Event-Event Conditional	Mixed Conjunctions	Positive	17%
	Both	Only When, Event-Event Conditional	Consistent Conjunctions	Positive	11%
	N/A	None of the Above	N/A	N/A	5%

Table 13: Proportion of survey responses for most natural-sounding policies across five scenarios.

Pattern	**Percent of Participants Who...**			
	Need Help	Want Help	Would Use Tool	Would Be Annoyed by Tool
Different Triggers, Same Behavior	41.1%	56.7%	65.6%	23.3%
Same Behaviors	43.6%	56.4%	73.1%	24.4%
Different Triggers, Different Behaviors	47.3%	61.5%	71.4%	22.0%
Trust	52.5%	60.0%	68.8%	25.0%
Loops	48.7%	65.4%	75.6%	25.6%
Privacy	59.2%	71.1%	75.0%	15.8%
Chain	55.3%	74.5%	78.7%	18.1%
Same Except No Condition	61.8%	73.0%	76.4%	23.6%
Opposite Behaviors	54.4%	72.2%	68.4%	31.6%
Un-Paired Automations	68.4%	77.6%	76.3%	25.0%
Same Triggers, Different Behavior & Conditions	65.9%	75.8%	89.0%	26.4%
Extended Behavior	64.6%	73.4%	69.6%	25.3%
Chains with Opposite Behaviors	59.3%	63.7%	70.3%	23.1%

Table 14: Percent of participants who need/want help identifying TAP anti-patterns, whether they would use a tool to help identify such anti-patterns, and whether a tool would be annoying. Anti-patterns are sorted in order of easiest to understand to hardest to understand, based on participants responses in Figure 3.

On the Recruitment of Company Developers for Security Studies: Results from a Qualitative Interview Study

Raphael Serafini
Ruhr University Bochum

Marco Gutfleisch
Ruhr University Bochum

Stefan Albert Horstmann
Ruhr University Bochum

Alena Naiakshina
Ruhr University Bochum

Abstract

To address the issue of participant recruitment for security developer studies, researchers proposed using freelance online platforms or recruiting computer science (CS) students as proxies. However, recent studies showed that company developers performed better than freelancers or CS students in security developer studies. Additionally, studies on factors influencing usable security and privacy in companies make recruiting professionals indispensable. Therefore, we investigated influential factors on the motivation of software developers regularly employed in companies to participate in security studies. We conducted 30 semi-structured interviews on their perceptions of study factors concerning study design, recruitment methods, and data collection. We found that the study duration, topic, monetary compensation, and trust are influential factors for participation in developer studies. However, participants were concerned about high effort and weak performance in security tasks. Based on our findings, we provide recruitment and study design recommendations for future security research with company developers.

1 Introduction

Recruiting professional software developers for usable security studies is an ongoing challenge. Due to the small population size, lack of time, spread out geographical locations, and high cost [2–4, 30, 32, 33, 53, 66], researchers are often struggling to recruit participants, especially for quantitative studies or studies involving software development tools [1,20,32,44,60]. Therefore, Kaur et al. [29] and Tahaei et al. [60] compared different online recruitment platforms and samples (freelancers vs. CS students) for security developer studies concerning participants' programming skills and security knowledge. They proposed to either recruit freelancers, CS students or to use specific crowdsourcing platforms along with pre-screening surveys.

However, while the recruitment of CS students or freelancers might be useful for the investigation of different study design parameters (e.g., security prompting) or specific research questions (e.g., "Does a new system design improves the state of the art?") [42], past research also noticed significant differences in the preferences and performance of students and professionals in security developer studies [17,42,66]. In a password-storage study, Naiakshina et al. [42] found that professional software developers employed in companies submitted significantly more secure solutions and chose better security mechanisms than CS students or freelancers.

Research on factors influencing usable security and privacy in companies, such as the company context, organizational processes, security culture, or the communication between security and privacy experts and software developers, makes recruiting professional software developers from the industry indispensable [7, 8, 24, 34]. Further, understanding the challenges software developers face after the introduction of new regulations for security and privacy affecting companies (e.g., California Consumer Privacy Act (CCPA) or the General Data Protection Regulation (GDPR)) makes the recruitment of company developers necessary [5, 34, 58].

While different approaches exist to recruit professional software developers employed in companies [24, 42, 55], it is unclear yet how and where developers prefer to be contacted and recruited for security developer studies. Do they prefer to be contacted by researchers or sign up for a mailing list to receive study invitations? Does it matter whether researchers come from academia or industry? Which type of study would they be willing to participate in (online, lab, or field study)? Is the length and type of study (interview, survey, programming task) relevant? While in some security studies, developers

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.
USENIX Symposium on Usable Privacy and Security (SOUPS) 2023.
August 6–8, 2023, Anaheim, CA, United States.

participated for free [3, 14], in others, they received monetary compensation [31, 42]. No consensus among researchers exists on the type and the amount of compensation to receive higher participation rates yet.

To provide first insights into the study perceptions of company developers, we conducted 30 semi-structured interviews with professional software developers employed in companies and investigated the following research questions:

- **RQ1:** How do study factors affect the motivation of company developers to participate in security studies?
- **RQ2:** How and where do company developers prefer to be contacted for participant recruitment?
- **RQ3:** Which concerns do company developers have with study data collection?

To the best of our knowledge, this is the first qualitative study investigating influential factors on the motivation of professional software developers employed in companies to participate in empirical research studies. We recruited company developers from former studies and participants who indicated not having previously participated in studies. With this sample, we explored the motivations of company developers to participate in studies instead of deterrent reasons for refraining from study participation at all. Our study findings suggested that company developers prefer to be actively recruited for academic research studies with flexible time slots and receive monetary compensation linear to the study duration. However, security tasks were perceived as specifically challenging. Company developers were concerned that security tasks might be more complex than general programming tasks and thus less predictable in the effort. Additionally, if studies are conducted in accordance with their company, developers might feel the pressure of performance tests. Based on our findings, we provide recommendations for future security studies with company developers.

2 Related Work

In this section, we discuss existing guidelines and recommendations for participant recruitment in software engineering studies and work on factors affecting participant motivation.

2.1 Guidelines on Participant Recruitment

Existing work on conducting software engineering studies assists researchers by providing them with guidelines and case studies on how to set up specific study tasks, such as surveys [41], interviews [26], or experiments [57, 63]. The recruitment advice found in the literature often focuses on recruiting a reliable and representative sample [48] or providing testimonials on specific issues and pitfalls during participant recruitment [19]. While guidelines exist on how to recruit professionals for software engineering studies, they focus on how to conduct studies methodologically [11], how to recruit the right participants [49], or how to establish cooperation between researchers and companies [52]. There exist also recruitment guidelines from different fields, such as health sciences [36], that discuss barriers to study participation. However, these are tied to a clinical context and might not be generalized to a software engineering context. While these recommendations are helpful for reliably collecting data from recruited participants, they provide little insight into how to motivate suitable individual software developers more efficiently and in significant numbers.

Thus, to conduct empirical studies with developers, researchers often relied on their industrial and personal contacts [24, 29, 42] or convenience samples such as CS students [9, 32, 44, 59, 60], crowdsourcing (e.g., Appen, Clickworker, MTurk or Prolific) and freelancer platforms (e.g., Upwork, Freelancer) for participant recruitment [16, 17, 23, 25, 29, 43]. These platforms come with challenges such as unreliable data [17, 47, 56], weak built-in tools [51], or contract limitations (e.g., participants can be compensated only via Upwork for the first two years) [23]. Therefore, Kaur et al. [29] and Tahaei et al. [60] compared different recruitment platforms and samples and provided recommendations on specific crowdsourcing and freelancer platforms for participant recruitment. However, it is unclear yet, how professional software developers employed in companies can be recruited sustainably. We investigated factors affecting company developers' willingness to participate in research studies specifically with a security focus and provide insights into viable recruitment strategies.

2.2 Studies on Participant Motivation

In [54], Smith et al. studied factors associated with recruiting professional developers for software engineering surveys. These factors were based on research of study design, persuasion, and the researchers' experience in conducting previous surveys. A post-hoc analysis of surveys from past research showed that scarcity cues (e.g., time limits or a maximum number of participants) and similarity cues (e.g., being part of the same company) increased the chance of a high survey response rate. Interestingly, besides monetary compensation, complimenting participants or using humor in the study invitation made survey invitations even more successful.

Based on their lessons learned from qualitative research with developers, Brandt et al. [13] provided recommendations on improving the recruitment process for security developer studies. First, they suggested motivating participants by convincing them that there is value beyond monetary compensation for participating in a study, such as learning about a new tool or security technique. Second, they proposed stating why an individual participant is valuable for the particular research study, e.g., competence in a specific domain. Third, they recommended reducing the effort for participants, e.g., by using a

calendar service for meetings or providing an online development environment instead of setting up an environment themselves. In addition, it might be beneficial to appeal to different types of motivation, such as fun, drive to produce knowledge, social connection, and self-improvement to increase study participation rates [6]. In the context of an end-user study, Hsieh et al. [27] investigated the influence of incentives on participation bias in surveys. They found that different incentives influenced the task outcome, i.e., different incentives attracted participants with different motivations. For example, charity rewards attracted participants who valued universalism and benevolence. Using diverse stimuli might increase the response rate and draw a more varied sample, as compensation is quite individual. Offering monetary compensation, however, was the most effective way to increase response rates. Compared to the previous studies, we conducted semi-structured interviews with professional software developers from companies. We explored their attitudes, expectations, and perceptions concerning recruitment strategies, challenges, and study factors such as the study type, security, and compensation.

3 Methodology

To investigate influential factors on company developers' willingness to participate in security studies, we conducted 30 semi-structured interviews with regularly employed professional software developers. All 30 interviews were conducted online by the same researcher using Zoom Video Communications [67]. While participants did not have to turn on their cameras mandatory, the audio was recorded using OBS Studio [46] and transcribed. At the end of the interviews, participants were asked to fill out a short survey on their demographics (Appendix B). In total, we collected 23.25 hours of interviews with a mean interview length of 46.5 minutes (min: 31, max: 64, median: 46).

3.1 Study Factors

We developed a list of relevant study factors based on related work and several meetings and discussions with one highly experienced (5 years of experience in conducting studies with end users and developers) and one senior researcher from the human-centric security research field (more than 20 years of experience). First, we were interested in developers' perceptions of different study design parameters. We asked participants whether the study topic might affect their decision to participate in a scientific study, especially focusing on security or involving security tasks. Additionally, we asked for reasons for their preferred study types: *Online, Lab, and Field Studies*, and study tasks: *Survey, Interview, and Practical Task*. Practical tasks might require writing programming code, a code review, or a protocol. We also asked participants how the study duration and compensation might influence their willingness to participate in a study. For example, whether they would be interested in participating in a workshop or receiving licenses for software products they might use for their job or free-time. Second, we investigated how and where participants prefer to be recruited by researchers. We asked for different recruitment channels such as email, company research cooperation, conferences, workshops, social networking platforms (e.g., Xing [65], LinkedIn [35], Facebook/Meta [38], Twitter [62]) and whether participants would prefer to be recruited by an active or passive recruitment strategy. Participants might receive personal study invitations from researchers or mailing lists if being actively recruited. With passive recruitment, participants would be required to search for study invitations, e.g., on the Web. We also wondered whether researchers' backgrounds, such as computer scientists or psychologists, might affect participants' willingness to participate in a study. Third, we explored their trust in researchers' organizations, such as academia or the industry. Since different data can be collected from participants in a study, we also explored their privacy concerns with data collection, storage, and usage.

After participants reported their experience with previous research studies, they were presented with the different study factors relevant to the context of study design, recruitment process, and trust in research:

- **Study design:** Study topic, study type, study task, study length, compensation
- **Recruitment:** Active/passive recruitment, recruitment channels, researcher background
- **Data collection:** Trust in researchers' organization, data collection, storage and usage

To test our study setting and interview guideline, we conducted a pilot study with two CS students and one professional software developer. Both CS students worked halftime in a company and provided valuable feedback on the interview guideline. We adapted the guideline by adding additional follow-up questions based on the participants' feedback. The final interview guideline can be found in Appendix C.

3.2 Participants

For participant recruitment, we used our personal contacts and a database of professional software developers who have already participated in past empirical research studies and agreed to be contacted for future studies. Thus, we ensured a broad spectrum of experience with different recruitment channels, study parameters, and study tasks with a security focus. Nineteen interested participants signed up for the study. Additionally, participants shared our study invitation with their friends and colleagues. Thus, 64 additional interested participants signed up for the study. To ensure only professional software developers regularly employed in companies would participate in our study, we asked all participants to fill out

Table 1: Demographics of the 30 participants

Gender	Male: 26, Female: 4
Age	min = 22, max = 53, sd = 9.13, median = 33, mean = 35
General Development Experience [years]	min = 4, max = 30, sd = 8.22, median = 12, mean = 14.2
Working in Company [years]	min = 2, max = 30, sd = 7.02, median = 7 , mean = 10
Weekly Work Hours	min = 20, max = 50, sd = 9.87, median = 40, mean = 37.3
No. of Employees in Company	1000+: 10, 500-999: 5, 250-499: 2, 10-249: 10, 1-9: 3
Company Type	Web Development: 12, Frameworks/Libraries: 4 Consulting: 4, Industry: 6, Other: 4
Company Security Focus	Yes: 18, No: 12
Experience with Security Tasks	Yes: 22, No: 8
Participated in Studies Before	Non-Security: 9, Security: 15, None: 5, NA: 1
No. of Studies Participated Before	min = 0, max = 10, sd = 2.34, median = 2, mean = 2.35, NA: 3

a pre-screening questionnaire (Appendix A). We excluded participants who indicated working less than part-time at a company or if software development was not part of their job. Out of 40 invited company developers, 30 participated in our study. We did not receive a response from the remaining ten participants concerning an interview appointment. Twenty-eight interviews were conducted in German and two in English. Participants received 100€ as compensation.

An overview of our participants' demographics can be found in Table 1. All participants were regularly employed as software developers in German companies. On average, participants were 35 years old. Twenty-six were male, and four were female. On average, participants worked for ten years in a software development company, with a mean of 37 hours per week. Our participants had an average of 14 years of experience with software development. Most participants worked in companies with 10-249 or at least 1000 employees. Twelve participants reported working in a company specialed in web development, while others were employed in consulting or industrial companies, as well as in companies specialed in the development of libraries, frameworks, or middleware. Eighteen participants indicated working for companies with a security focus. Twenty-two of the 30 participants were experienced with working on security-related tasks in their company. Twenty-four participants stated to have already participated in a scientific study in the past, of whom 15 have already participated in a study with a security focus.

3.3 Evaluation

We analyzed the transcribed interviews with the software MAXQDA [64] by using thematic analysis [12]. After two researchers (R1 and R2) coded the first three interviews individually, they agreed on one codebook through discussion. Based on this codebook, all the remaining interviews were analyzed by the first two researchers, R1 and R2, and a third researcher, R3. Researcher R1 coded all 30 interviews. R2 coded a random set of 15 and R3 the remaining 15 interviews. As suggested by McDonald et al. [37], we calculated an inter-coder agreement to ensure the consistency of the code application. It was calculated using Cohen's kappa coefficient (κ) [15]. The inter-coder agreement between R1 and R2 measured 0.77, and R1 and R3 0.78. A value above 0.75 is considered to be a high level of coding agreement [21]. The codebook can be found in Appendix D. Reported quotes were translated using DeepL [18] and adapted by the researchers in necessary cases to convey the quotes' meaning properly.

3.4 Limitations

This study has several limitations that need to be considered when interpreting the results. First, we cannot claim the completeness of the presented list of different study factors relevant in the context of study design and participant recruitment. There may be other factors influencing the motivation of company developers to participate in empirical research studies that are not covered in this study. However, we provide preliminary findings on factors that might be worth to be further explored in future research. Second, since we asked about factors influencing developers' decision to participate in a research study, results may suffer from a social-desirability bias. In addition, since our participants were willing to participate in research studies, they might be favorable to research and thus might have other views than those who abstained from this or study participation at all. As such, we can only make claims of motivations and preferences of participants willing to participate in research studies. Third, we recruited professional software developers regularly employed in German companies. Our study findings might not apply to developers from other parts of the world. Further work is needed to make any generalizable statements.

3.5 Ethics

Our institution does not have a formal Institutional Review Board (IRB) process for computer science studies, but the study protocol was cleared with the institutional data protection officer. Our participants were provided with a consent

form complying with the General Data Protection Regulations (GDPR). They were informed about the practices used to process and store their data and that they could withdraw their data during or after the study without any consequences. The audio recordings were deleted after transcription. We assured all participants they would be informed about the study results, and only anonymized data would be published.

4 Results

In this section, we present the motivational factors and barriers concerning study design, participant recruitment, and data collection we identified in the interviews. An overview of the factors can be found in Table 2. To report statements, we labeled participants P1-P30. While we note how many participants stated specific themes to indicate their frequency and distribution, we do not aim to generate quantitative results.

4.1 Study Design

In the following, we present relevant insights into the study design of security studies with software developers concerning the study topic, study type, task, length, and compensation.

4.1.1 Study Topic

Our participants frequently noted an interest in the study topic as motivation for their participation. They often perceived the topic as an opportunity to learn something useful or to work on something in line with their interest: *"Sure, the topic has to interest me somehow"* — [P19]. Six participants (e.g., P4, P7, P10, P12, P24) were willing to receive less compensation if they could work on a study topic that is in line with their motivation to participate: *"If it is a topic that interests me [...] I would also participate in studies, but that is not where I earn my money"* — [P4]. However, the opposite was also true. Participants claimed they would need a higher compensation to compensate for their disinterest in the study: *"So, if the topic does not interest me at all, then you could lure someone or me with money"* — [P2]. In addition, participants often mentioned they were less likely to participate if they were unsure about fulfilling the requirements for study participation. They worried they could negatively influence the study results and thus told to be hesitant to participate: *"The study topic, yes, I think it is essential that I [...] have to know, if I can say something about it or not"* — [P4].

We did not observe differences in the statements of participants who have already participated in security studies (15/30) and those who participated in software engineering studies (9/30) or did not participate in studies at all (5/30). However, participants often explicitly referred to IT security as a topic they would consider when participating in a scientific study: *"This is a topic that also interests me in a professional context"* — [P16]. One participant stated that, especially in the context of IT security, they can *"[...] learn (something themselves)"* — [P10]. Participants felt that developer studies with a focus on security might be a good opportunity to update their knowledge concerning state-of-the-art security: *"If you're programming something, anything security-related, and you realize that I haven't updated my knowledge in the last ten years, that kind of added value"* — [P18]. Two participants also stated that they liked to be challenged: *"It has a playful component. You challenge yourself. I can see that I'm doing a good job"* — [P29]. The motivation to learn something new is also reflected in the non-monetary compensation our participants suggested. For instance, attending a workshop: *"[Offer] an information event on the IT security topic there, that would [...] make that more interesting again"* — [P16] or receiving a security certificate: *"So something like a security certification that you could maybe then obtain"* — [P10]. However, P11 stated that it is hard to estimate the difficulty of security-related tasks. Therefore, they would like to be informed about the extent of effort the task might need: *"Especially in the area of security, it is always good to know in advance at what level such a study will ultimately be conducted, how deep the whole thing goes technically. In the field of software development, it is always assumed that things will become technical very quickly"* — [P11].

4.1.2 Study Type

All but one participant stated online studies to be their preferred study type. As advantages, participants mentioned the flexibility and duration of an online study: *"Because the time required is less than for another study"* — [P1], comfort: *"You are at home in your comfort zone"* — [P28] and that they do not have to travel in the COVID-19 pandemic: *"I do not have to leave the house. [...] Right now also with all that Corona time. This is just more pleasant [...]"* — [P26]. Eighteen participants (e.g., P1, P5, P11, P17, P30) mentioned that they expected less compensation in an online study compared to a field or laboratory study.

All participants stated that participating in a lab study might be more challenging since they would need to travel. Twenty-one participants (e.g., P2, P5, P14, P21, P26) noted that they would expect additional compensation for traveling: *"If that is now somehow locally at the university, where I had to drive there, then the travel costs would have to be somehow considered, plus the time expenditure [...]"* — [P2]. P3 described that they would have taken a day off from work to participate in a laboratory setting: *"The planning, that simply my life gets affected more, that I have to organize more, maybe take a day off or something like that"* — [P3]. Our participants had different opinions on field studies. Twelve participants (e.g., P3, P6, P9, P22, P27) were open-minded toward field studies, provided their employer would previously approve the study participation. Others (e.g., P3, P11, P12, P26, P28) explained

Table 2: Motivational factors and barriers concerning study design, recruitment, and data collection

Study Factor	Factor Level	Motivation	Barrier
Study Topic	**Software**	In line with developer hobby or job	Unexciting topic
	Engineering	Self-improvement	Uncertainty about requirements
		Learning effect	Uncertainty about effort
	Security	Update knowledge	Security performance test
		Security certificate	
Study Type		Flexibility	
	Online	Working from home/while traveling	
		Anonymity	
			Traveling time and expenses
	Laboratory	Social interaction	Issues due to Covid-19
			Requirement to take a day off work
			Concerns about data privacy of clients
	Field	Combining work with study	Conflict when working with clients
			Conflict when working from home
			Company reputation in case of weak security performance
Study Task		Short	
		Easy to do	
	Survey	Flexibility	Length
		Anonymity	
		Variety to job	
		Social interaction	
	Interview	Variety to job	Length
		Flexibility	
		Learning effect	Uncertainty about task requirements
	Practical	Personal challenge	Uncertainty about task effort
			Length
Study Length		Control of time scheduling	Length
		Time flexibility	Rigid date scheduling and loss of flexibility
		Continuous compensation in long-term studies	Long-term commitment
Compensation		Monetary compensation	Skepticism toward too high compensation
Recruitment	**Strategy**	Active recruitment	Passive recruitment
		Forwarded by friends and colleagues	Spam
	Company	Trust in approved study	Lack of time
			Performance test
		Asynchronous nature of email	Headhunting on networking platforms
	Channels	Unsolicited emailing with verifiable identity	Unsolicited emailing without verifiable identity
		Study advertisement at workshop or conference	
Data Collection		Trust in academia	Distrust in industry
		GDPR compliance	Sharing sensitive data

that they have been in home office for a long time and thus did not believe conducting a field study would be a good idea: *"But when it comes to seeing myself in my natural environment, for example, I don't know if the field study isn't the best way to go, in quotes, because this home office situation isn't necessarily what you prefer"* — [P6]. P11 would also like to avoid involving customers: *"[. . .] I am also visiting customers as a software developer, and I do not need someone there who conducts a study, who stands behind me, and who looks over my shoulder"* — [P11].

4.1.3 Study Task

We asked participants about taking part in a study involving a practical task. Some indicated they would appreciate taking part in practical experiments and would prefer it over an interview or survey: *"Yeah, because in practical tasks, [. . .] you will have the opportunity to learn something"* — [P21]. Others (11 participants, e.g., P4, P16, P18, P22, P25) mentioned uncertainty about practical tasks, as they do not know what to expect, how long they will need, and whether they might fit the task requirements: *"[. . .] but then the framework conditions would have to be a bit clearer in advance, and basically I would also have to know what kind of tasks I would be facing. So the uncertainty would then rather strengthen the aversion"* — [P16]. Further, P6 described that in programming tasks, they had to engage deeper with the topic: *"I associate a practical task with the fact that I have to get involved with the subject. That means that in my mind, it's always associated with more effort"* — [P6]. Almost all participants perceived a practical task as more difficult in general. When asked if the study task influences the compensation they expect, half of

the participants wanted to receive higher compensation for a practical task compared to a survey or an interview: *"I would [...] expect a higher compensation for the practical work, definitely"* — [P24]. Further, P16 compared the work on a practical task with an *"examination situation"* — [P16].

Participants often preferred surveys and interviews over practical tasks, as they perceived them as a variety of their working tasks: *"I am coding, doing it the entire day. And then an interview is a welcome diversion"* — [P2]. Social interactions were frequently referred to as a positive factor associated with interviews: *"Because then I would rather interact, instead of standing alone, in front of some piece of paper, or sitting and having the feeling, oh, I still have to fill out this piece of paper"* — [P6]. Further, participants valued the flexibility surveys offer besides lower attention required: *"And because I do not have to expend so much effort and because I don't have to think much, you can do that on the side"* — [P7]. Thus, participants accepted less compensation for surveys: *"So, with such a click survey, I would not necessarily expect such an amount as a reward for the interview here. Because you can do it in between, while you are doing something else, or you can interrupt it"* — [P16]. However, participants disliked long surveys due to boredom and mental fatigue: *"But whenever I am filling out a survey, especially if it is longer, I get to the point where I drift in thought and simply start clicking through, to finish"* — [P8].

4.1.4 Study Length

Many participants considered a long duration as a barrier to taking part in a scientific study, which can not be broken through other forms of motivation: *"One can then simply not participate.[...] But I think that [this] is not a factor now for better hourly payment"* — [P17]. Fifteen participants stated that their time is limited, e.g., because of family responsibilities: *"I have a job and a family, so it is always difficult to find the time slot [...]"* — [P16], or the need for recreation: *"Although I am employed full time, [...] I should also relieve my head a bit sometimes"* — [P17]. Participants were also asked about their willingness to participate in long-term studies spanning months or even years.

Some participants stated that while they might be interested in participating, they would not like to agree on fixed dates, weeks, or even months ahead: *"I couldn't say, okay, I can set this exact block for myself every two weeks or every month"* — [P18]. P19 stated that it would be beneficial if you can *"[...] determine the time frame accordingly, [so] that you can also do something [...] on the weekend or [...] in the evening after work"* — [P19]. Another participant liked the idea of a long-term study: *"It would be super cool if such a study is a long-term study and you get money again and again"* — [P5]. However, 15 participants (e.g., P4, P7, P17, P22, P29) preferred a study length of one to four hours.

Ten participants (e.g., P4, P17, P19, P25, P28) equated time to money and compared the study compensation with their salary: *"[...] Either I sit for two hours or more in a study or I just work longer, then I think most people start to wonder whether it's worth it"* — [P18]. However, 18 participants (e.g., P2, P8, P11, P16, P28) did not expect a higher hourly wage for more extended studies. Instead, they expected linear payment: *"So basically that you have some kind of hourly rate that you have for an interview, if you say one hour one hundred euros, if it's two hours it's two hundred euros"* — [P16]. If taking a vacation day for a study with a long time frame would be necessary, six participants (e.g., P5, P11, P19, P24, P25) expected to be compensated for the loss of a working day: *"So then it would have to be much better paid than the day off costs me [...]"* — [P24]. In a multi-day time frame, 15 participants (e.g., P5, P9, P11, P19, P27) stated to *"[...] rather not participate if [it would require a] day or something, then the will is not so high"* — [P16]. Overall, many participants stated that flexibility is essential. Having control over the time frame might increase the willingness to participate even in a long-time framed study: *"So the more flexible you are in choosing the time slot, the longer you can manage the hours for a study"* — [P2].

4.1.5 Compensation

Almost all participants (28/30) preferred receiving monetary compensation: *"Are there serious compensation suggestions other than money?"* — [P15]. Nine Participants perceived compensation methods such as Amazon vouchers as appropriate as well: *"Amazon is like cash to me actually"* — [P2]. Fourteen (e.g., P5, P10, P16, P21, P26) opted for participation in a workshop as compensation. Still, restrictions were mentioned for workshops: *"[...] especially for such an on-site study, if it is somehow integrated into a workshop, where I then take something away for myself, it would be interesting for me"* — [P16]. As a further barrier to workshop participation, participants stated that *"with training offers [...] it's [...] again associated with having to sort of balance out when, how, where"* — [P6]. They would need to coordinate with their employer first to get some time off: *"So if, for example, some conference is to take place now, [then] I would have to ask my employer[...] and they would have to spend the money accordingly. And that's exactly what you could save at that point"* — [P26]. Further, participants mentioned that they were concerned a workshop might force them to take part at a specific date, which might limit their flexibility: *"[...] Then I probably lack the time flexibility again"* — [P5].

Eighteen participants were willing to accept a software license as compensation for study participation if it fitted their requirements: *" [...] only [...] if it is something I am working on right now"* — [P5]. One participant explained that even if they would be interested in getting a license, they would still compare this to the actual monetary value of the product: *"[...] nevertheless, I would weigh afterward again*

[. . .], what is the financially equivalent value of it [. . .]" — [P1].

Most participants (24/30) stated that a higher compensation would increase their willingness to participate in a study, and thus they might ignore the challenges they perceive. P3 specifically said that they would also do *"[. . .] unpleasant things"* — [P3] if the amount of compensation would be appropriate. Many participants reported that there exists a hard limit for study participation. For example, some stated they must fulfill their family and work responsibilities. Therefore, taking part in studies that last more than some hours might not be possible - regardless of the compensation amount: *"[. . .] because I have a family and two children. So I have to find a place to fit it all in. That's why the time factor is also important in any case"* — [P6]. Five participants also mentioned that they might be skeptical if the compensation would be too high: *"I would be more concerned if a private company paid me too much"* — [P3]. Interestingly, 100 euros for an interview study with software developers was perceived as too high by P14: *"So I definitely think it's way too much in this study"* — [P14].

4.1.6 Other Factors

Beyond monetary compensation, other factors were mentioned developers perceived as a motivation to participate in scientific research. Some participants wanted to contribute to society and maintain *"[. . .] the dialogue, between the generations, between old and young, experienced and inexperienced"* — [P4] because it might be valuable for *"the community"* — [P13]. Thirteen participants (e.g., P3, P8, P10, P19, P26) stated that they were motivated to help since they were curious *"what will happen afterward (with the results)"* — [P4]. Others were motivated by recognizing their own mistakes and receiving feedback on their work: *"if they will give me some feedback"* — [P23]. Some liked the idea of receiving *"an appreciation letter or basically certificate"* — [P21] as recognition for their efforts, and others felt appreciated *"[. . .] to be asked as an expert about something"* — [P5]. Twenty participants (e.g., P6, P9, P12, P19, P26) also reported that their work activity is in line with their preferences in their freetime, and thus they might consider participating in a scientific study: *"My profession is also my hobby"* — [P7].

4.2 Recruitment

When asked what should be included in the study invitation, 23 participants (e.g., P1, P9, P10, P16, P24) stated that *"the theme must fit me in any case [. . .]"* — [P26]. Another participant stated that they would want to be personally addressed because *"it's just much more personal, and then you also have a good feeling that you're helping (with) something"* — [P5]. For short studies, some participants preferred to start with the task after receiving the study invitation immediately: *"If one [. . .] would have to sign the consent form or make an appointment or something similar, then rather not"* — [P2].

4.2.1 Active & Passive Recruitment

Twenty-five participants (e.g., P2, P4, P12, P17, P25) favored active over passive recruitment. One of the disadvantages of passive recruitment most participants faced was *"not knowing where to find it"* — [P28]. Seventeen participants (e.g., P1, P3, P12, P22, P29) considered not having to look for a study themselves as an advantage of active recruitment, as summarized by one of the participants: *"For me, this is just right. I am a lazy person"* — [P25]. Overall, many participants considered being personally addressed in active recruitment as beneficial since *"that just comes much more personal"* — [P5] and they knew *"that I was not simply addressed randomly"* — [P4], but rather for the skill they possess. Not all attitudes towards active recruitment were positive. Seven participants (e.g., P9, P15, P22, P27, P29) worried that active recruitment could result in spam: *"If twenty emails per week are coming in now, I guess it's time for me to activate the spam folder"* — [P29]. Most participants (21/30) favored asynchronous communication, especially for first contact, which applies to both types of recruitment: *"I like all forms of first contact where I don't have to react spontaneously, whether it's an email or an ad"* — [P3]. Nine participants also had positive attitudes towards study invitations being forwarded by friends or colleagues since they have *"[. . .] a few bonus points of trust on top"* — [P9], especially if they already received their study compensation.

4.2.2 Researcher Background

Researchers' background was often linked to the study topic. For instance, 14 participants (e.g., P2, P8, P10, P12, P30) stated their decision to participate in a study would not be affected by involving researchers from fields other than computer science: *"This would not be so important to me"* — [P19]. However, P2 noted that they would be more willing to help researchers if they shared the same background: *"When I see that is also a computer scientist, then I still tend to be more helpful"* — [P2]. Three participants (P15, P16, P17) considered a researcher's background as essential for their decision to participate in a research study: *"I think you have to have a background there"* — [P16]. Nine participants (e.g., P1, P3, P13, P14, P21) stated that *"it really depends on the topic, if it is supposed to be more interactive, the IT person will probably have more understanding of the terms"* — [P9]. P3 stated that *"Experience would be more critical"* — [P3].

4.2.3 Company

Ten participants (e.g., P3, P12, P25, P26, P29) believed their employer would be willing to share a study invitation among their employees. However, *"when there is a lack of time in*

the company, [the employer would] say: "Ah, we'd better not pass that on"" — [P1] even though they promised researchers otherwise. Other concerns had been that *"in the worst case, the employer is in cahoots with the researcher and uses the situation to test my performance somehow"* — [P29] or that they might be forced to participate by their employer: *"That would be fine, too, if the employer didn't push me to do it"* — [P17]. Four participants (P2, P9, P11, P25) complained about disguised headhunters: *"You wouldn't get through to me at all, very quickly put a stop to it and say, yes, that's actually just a hidden inquiry. In reality, you are a recruiter"* — [P11]. Still, study invitations approved by an employer were trusted more: *"I mean, if the employer has already run this, certain checks have already been run in advance, which means that I can really assume that it is a serious study"* — [P20].

4.2.4 Recruitment Channels

Almost all participants preferred email as their primary recruitment channel: *"So as a computer scientist, I have to be honest, I'm most comfortable with the email form"* — [P17], due to its asynchronous nature: *"Because of the asynchronous and because it is then just a bit personal"* — [P6]. Unsolicited emailing, participants accepted under certain conditions: *"I would like to know where they got my email address from"* — [P30] and who the sender is, since *"if it was in my inbox from someone I didn't know, I would be skeptical. That's when the alarm bells go on"* — [P29]. Most participants indicated using social networking sites such as LinkedIn or Xing for their professional contacts. Some participants agreed on receiving invitations through these channels: *"I think, so something like Xing or LinkedIn or something like that, I think you would expect it anyway"* — [P18]. Others (eight participants, e.g., P6, P14, P26, P28, P30) refused to be contacted through such channels: *"[. . .] because I personally perceive it less as a place for news"* — [P6]. Some participants (P9, P10, P24, P28) stated that they were exhausted by the mass of recruiters contacting them. Thus, they might end up skipping legitimate study invitations: *"I usually blindly reject all requests with this automatic no-thank-you answer. A request like that could certainly be overseen"* — [P24].

Participants had mixed feelings about being recruited at a conference or a workshop. One participant stated that since they are already at the venue, they might participate in the study: *"I was at a conference where this was asked, where I participated in a small survey, especially if you are on site anyway, the hurdle is of course very low if you can just do it spontaneously"* — [P3]. Some might be more willing to participate if a speaker shares a study invitation with the audience while keeping it asynchronous. One participant stated: *"I'm generally not the one who likes to be addressed by booths or something, but if it's a conference and the speaker says, well, pass on the information, or it's also advertised somewhere, whatever"* — [P10].

Most participants indicated rarely using social networking platforms such as Facebook or Twitter. One participant questioned how serious study invitations, shared via these channels, are: *"I don't really think it's serious either. Honestly, I wouldn't believe it's anything real either"* — [P18]. Still, one participant stated that the channel is not essential, but rather who is recruiting and how participants were addressed: *"So basically, I wouldn't say if you brought it to my attention [. . .] via Twitter, I wouldn't do that because that came via Twitter"* — [P6].

4.3 Data Collection

Participants preferred data collection to be fast and easy: *"So, if you work with modern tools, for example, GitHub, where you can upload and download the code, that would have been more comfortable. In that case, I have, I think, sent it via email. I am not entirely sure, but it was a bit of back and forth"* — [P2]. Some participants were irritated by the collection of the same data multiple times: *"But questions which were already asked, then discussed afterward again, I found a bit odd"* — [P11].

4.3.1 Researcher Organization

Participants had different attitudes toward the type of organization conducting the study. Many participants mentioned they had a high level of trust toward public institutions like universities. They thus were more comfortable with data sharing compared to private corporations: *"Because I am more prepared to participate in public studies and use my personal resources, and my data is part of that, compared to private studies"* — [P12]. In addition, some participants believed that universities had a lower budget compared to corporate entities and thus expected a lower compensation from studies done by universities: *"To be honest, I thought that the university pays worse. [. . .] I think [companies] also have more money"* — [P18]. Further, some participants raised concerns regarding illegal data collection by companies during studies, something they assumed was less likely to occur with public institutions: *"So, there are some things, like market research institutes, where you can install some kind of client, [. . .] which then collects all kind of data in the depth of windows and sends to them, where you have no control in the first place"* — [P10]. Some participants assumed German companies were more likely to collect their data in compliance with GDPR.

4.3.2 Privacy Concerns

Some participants felt that a high compensation for their data would raise suspicion about the usage of their data: *"If the compensation for things like that is higher, I (would) have concerns about whether [. . .] (it is) too high and why"* — [P3].

In many cases, participants preferred to share as little data as possible. However, 16 participants (e.g., P6, P11, P15, P21, P30) argued they would be willing to share their data if the usage was in line with the research topic: *"Eye movement, well, I really do not know, [...] I would have to be sure how it is processed"* — [P15]. Questions concerning sensitive data raised the issue of anonymity for many participants: *"Well, to be honest, there really was [that] moment in a survey after a study. [...] I would say that not everyone honestly answers such sensitive questions because, in the end, you are still not sure how anonymous everything is. So there is a certain amount of trust, but still"* — [P17]. This was linked to the study type, as participants frequently mentioned the increased anonymity of online studies as a positive factor: *"This is the most significant advantage for me. Maybe also that it is a little anonymous"* — [P7]. By contrast, many participants were skeptical about field studies, as data collected at their workplace might be critical: *"And then this person asked me about security-relevant features, about knowledge, that you have to have for it and finds out, I do not know, how to solve this problem best. And you could use that against the company where I work"* — [P5].

5 Discussion

In this section, we discuss our research questions. We provide insights into recruitment obstacles specifically related to the company context. Past research with students showed that they were often available for laboratory studies [10, 32, 61], flexible in time [40], and did not worry about the consequences of their study performance [43]. Besides logistical issues, our participants reported being less flexible in time and worried about their and their company's reputation, unpredictable effort, and potentially bad performance communicated to their company. Since most of our participants were already experienced in study participation, preferences on the indicated factors for specific types of studies we identified might only apply to participants willing to participate in research studies. While we cannot make any claims about those who opted not to participate in this study or studies in general, our results indicated preferences for different types of studies on recurring participants. As such, our results can help researchers retain interested participants for various types of studies.

RQ1: Influence of study factors on study participation. Our participants indicated monetary compensation as the most influential factor motivating them to participate in research studies. Participants felt that high rewards could compensate for working on research topics they might not be highly interested in or participating in long-time framed studies. They also expected higher rewards if working on practical security tasks or required to take a day off from their regular work (e.g., traveling for a lab study). However, they were willing to work on study tasks in their free time as long as they did not conflict with their family responsibilities. Thus, flexibility and schedule control was essential for accepting study invitations, which were not considered to be compensated by higher monetary compensation.

Our participants generally did not expect a higher payment for long-time framed studies but rather a linear hourly rate. Still, they were skeptical about high payment if they could not relate it to the expected effort. Thus, participants expected less payment for online studies, short surveys, or studies conducted by academic researchers in a university context. While participants perceived vouchers (e.g., Amazon) as another form of monetary compensation, workshop participation or software licenses might be less promising forms of compensation due to highly individual demands. Besides compensation, participants were often motivated by idealism to contribute to society, self-improvement, or receive feedback on their work.

With the continuous threat of security vulnerabilities attackers might exploit, it seems participants felt that developer studies focusing on security might be an excellent opportunity to evolve and update their security skills and knowledge concerning recent security topics. They indicated that security tasks would have a positive learning effect. Participants would also appreciate receiving a security certificate after study participation. Especially, practical tasks might be an excellent way to improve their security skills, and thus participants would be willing to participate in security developer studies. However, participants were concerned that security tasks might require more effort than expected. They were also often unsure whether they fulfilled the requirements. Therefore, they would like to have clear information on the task requirements. Additionally, they worried about weak performance in security tasks, which might be linked to their company (e.g., in a field study).

RQ2: Recruitment. With active recruitment by researchers, developers were not required to search for study invitations, which might save time and effort. Additionally, participants indicated being unsure where to look for study advertisements. They favored email as a recruitment channel since it allows asynchronous communication. Further, they appreciated the personal contact and assurance they were recruited based on their competence in the studied field. However, participants were concerned about unsolicited emailing or the uncertainty about where their contact information was gathered from. Therefore, study invitations forwarded by friends and colleagues - especially if they had already received their study compensation - were trusted the most and thus might increase the likelihood of positive responses. Participants trusted study invitations distributed over their company since they felt their employer had already approved them. However, they worried about performance tests and the pressure of being forced to participate in the study. While social networking platforms were considered suitable for participant recruit-

ment in a professional context (e.g., LinkedIn, Xing), some participants were unsure how serious invitations were from social networking platforms in a more personal context (e.g., Facebook/Meta, Twitter). However, study advertisements at workshops or conferences were perceived as a suitable way to recruit participants.

RQ3: Data collection. Compared to the industry, research institutions like universities enjoy great trust among company developers, which is why most developers indicated being more willing to share their data. Participants preferred to share as little data as possible unless they were informed about data usage. They also indicated that the data collected should be in line with the research topic and comprehensible. However, security studies involving very high compensation raised concerns about data usage. Surveys and online studies were considered more anonymous than other study types and tasks. Thus, participants may be more willing to share personal data. Still, participants had no concerns about sharing source code as long as they were assured that it would not be used for commercial purposes. Data collection becomes even more vital when conducting field studies in a company because developers are worried about client information leakage or weak performance in security studies damaging their company's reputation. Overall, concerns about data collection might be mitigated by adequately informing participants about the data collection process and measures taken to ensure anonymity.

6 Recommendations

In this section, we provide recommendations for future usable security and privacy research studies based on our findings.

6.1 Study Design

In the following, we present our recommendations concerning study design.

Study Topic. For security studies with company developers, specific requirements and the expected effort should be made clear since uncertainty about tasks or topics might discourage developers from participating. Our participants were often unsure about the target group of study invitations or assumed other participants might be more suitable for the study topic. This is especially true for IT security since tasks in this domain are expected to be more challenging. If a study includes researcher-participant interaction, it might be beneficial to state a researcher's experience with the study topic.

Study Type. Online studies might be the preferred form for study participation for all study types (interview, survey, practical task) if high numbers of participants are required. If a laboratory setting has to be chosen, a virtual study environment, as proposed by Huaman et al. [28], might be a good option. Otherwise, explicitly compensating participants' time and travel expenses might be required.

Study Task. Many developers felt pressure on the performance of their practical tasks. High dropout rates due to potential task difficulty were also observed in past research, such as 42% in a study conducted by Acar et al. [1]. Thus, in usable security developer studies, participants might be made aware that the usability of a system but not their performance is tested. Concerns regarding task difficulty can also be alleviated by clearly stating the study requirements. Including an example task, as done in previous studies, can provide reassurance to participants [50, 51]. For surveys, developers preferred to start immediately after receiving the study invitation. Thus, including the survey link in the study invitation might be beneficial. This could be combined with checking for inclusion criteria and screening questions to ensure data quality. Due to the COVID-19 pandemic, participants reported having fewer social interactions. An interview can be a welcomed variety to company developers' daily jobs. For interviews, developers preferred minimal organizational effort. Using an easy scheduling service can reduce developers' effort and increase flexibility.

Study Length. Time flexibility is an essential factor for developers. Naiakshina et al. [42] reported that "developers dropped out because of a lack of time and [not managing] to solve the task in a functional way." Thus, long-time framed studies need also to be flexible in terms of time schedule, which might include offering dates on the weekends or outside office hours. It might also be beneficial to highlight the compensation, if available, associated with long-time framed studies.

Compensation. Monetary compensation is the most influential factor in increasing participant rates. However, participants also appreciated security certificates when participating in security developer studies. Thus, offering workshops after study participation, where participants can receive security certificates, might be a promising approach. In long-time framed studies, we suggest considering the hourly wages of company developers since they tend to compare the time invested in a study with their working time. Letting participants choose between different types of compensation besides intrinsic motivation (e.g., learning something new, updating knowledge) might improve response rates as well as sample quality as proposed by Hsieh et al. [27] in an end-user context. Interestingly, our participants considered the compensation of 100€ for a one-hour interview study rather high. While this can be the first indication for interview study payment, more research is needed on the payment levels for different study types, tasks, and lengths.

6.2 Recruitment

In the following, we present our recommendations concerning participant recruitment.

Recruitment Strategy. The study invitation should reflect and appreciate participants' competencies. It might be essential to explicitly state why their participation is vital for the study by acknowledging their skills and study topic requirements. In addition, participants might be informed how their participation can benefit themselves, the community, or society. For example, the conducted research might be helpful for the following generation, personal improvement, or specifically in the IT-security context, resolving security misconceptions and reducing security vulnerabilities. Previous work indicated that similarity cues improved response rates in surveys [54]. Thus, highlighting a shared professional or academic background might make a study invitation more appealing. Interestingly, all participants provided their consent to be contacted for future studies. They preferred this recruitment strategy due to the low effort. Therefore, we recommend asking participants for their consent to receive future study invitations after study participation. However, it would be helpful to avoid sending unsuitable study invitations and instead address the required target group of the study. For future work, it might be beneficial if a central database of interested software developers might be available to the research community. Alternatively, a central web page collecting advertisements for different research studies comparable to job openings might be a good form of informing interested participants.

Recruitment Channel. Email is the most preferred channel of communication as well as first contact because of its asynchronous nature and having everything in one place. Despite the common usage of social media and online forums for participant recruitment in security studies [7, 39, 45], most of our participants were unwilling to be recruited through these channels. Thus, we recommend using additional channels, as done in other security studies [29], alongside email. Research conducted by institutions like a university, research cooperation with companies, and study recommendations by friends and colleagues might increase trust in the study and thus participation rates. For unsolicited emailing, researchers might include where and how they extracted the contact information and explain how to remove the participant's contact information from this source. In addition, researchers might provide a verifiable way to check their identity or institution.

6.3 Data Collection

In the following, we present our recommendations concerning study data collection.

Data Collection, Storage and Usage. Explaining the purpose of re-iterating questions in surveys and interviews for data validity reasons might avoid irritating participants when the same information is gathered multiple times. In addition, data collection should be easy and relate to the study domain. Using internet hosting services such as GitHub for programming code submissions might be beneficial since developers indicated familiarity with them. If it is necessary to collect sensitive data from participants, it might be explained why this information is required. Participants mentioned fear of repercussions from the deanonymization of study data, a concern which was raised in past research as well [22]. Addressing and explaining the precautions taken may alleviate concerns.

Consent Form. Some participants were concerned with data collection, storage, and usage. Thus, we recommend presenting the consent form separately from the study invitation text to avoid overwhelming participants. Since our participants often referred to GDPR, researchers might inform participants whether data collection is in line with GDPR or CCPA, including third-party services used for data collection, storage, and analysis, e.g., transcription and hosting surveys.

7 Conclusion

Researchers often struggled to recruit developers for security studies. While using online freelancer platforms or inviting students can often be a good choice for participant recruitment, past research showed that the behavior of developers, freelancers, and CS students might differ in security studies. Additionally, usable security and privacy research in the company context requires recruiting professional software developers from the industry. Therefore, we conducted 30 semi-structured interviews with company developers to investigate influential factors on their willingness to participate in security developer studies. We found that participants mostly preferred to participate in short studies conducted online and in line with their interests. Developer studies with a security focus were perceived as specifically attractive since participants felt they might update their security knowledge and test their skills by receiving feedback on their performance. However, participants were concerned by the effort and weak performance effects security tasks might entail. Still, most barriers concerning study design, participant recruitment, and data collection might be countered by monetary compensation. Overall, company developers preferred to participate in studies with low effort but were willing to accept long-time frames if flexibility was ensured.

While offering first insights into influential factors on company developers' motivation to participate in security studies, further research is required with company developers from other parts of the world. Further, our findings suggested a strong influence of monetary compensation. Since no consensus between researchers exists, more research is needed in the context of different payment levels. Future studies might mitigate potential selection bias by conducting, e.g., large-scale and anonymous surveys to solicit participants without prior study experience. Since most of our participants stated, they are rarely invited to studies, utilizing different recruitment strategies might reach more first-time participants. However, more research is required to investigate barriers to study participation.

Acknowledgments

We thank our anonymous reviewers and shepherd for helping us improve our paper. This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC 2092 CASA - 390781972.

References

[1] Yasemin Acar, Michael Backes, Sascha Fahl, Simson Garfinkel, Doowon Kim, Michelle L. Mazurek, and Christian Stransky. Comparing the Usability of Cryptographic APIs. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 154–171, 2017.

[2] Yasemin Acar, Michael Backes, Sascha Fahl, Doowon Kim, Michelle L. Mazurek, and Christian Stransky. You Get Where You're Looking for: The Impact of Information Sources on Code Security. In *2016 IEEE Symposium on Security and Privacy (SP)*, pages 289–305, 2016.

[3] Yasemin Acar, Sascha Fahl, and Michelle L. Mazurek. You are Not Your Developer, Either: A Research Agenda for Usable Security and Privacy Research Beyond End Users. In *2016 IEEE Cybersecurity Development (SecDev)*, pages 3–8, 2016.

[4] Yasemin Acar, Christian Stransky, Dominik Wermke, Michelle L. Mazurek, and Sascha Fahl. Security Developer Studies with Github Users: Exploring a Convenience Sample. In *Proceedings of the Thirteenth USENIX Conference on Usable Privacy and Security*, SOUPS '17, page 81–95, USA, 2017. USENIX Association.

[5] Abdulrahman Alhazmi and Nalin Asanka Gamagedara Arachchilage. I'm All Ears! Listening to Software Developers on Putting GDPR Principles into Software Development Practice. *Personal Ubiquitous Comput.*, 25(5):879–892, oct 2021.

[6] Judd Antin and Aaron Shaw. Social Desirability Bias and Self-Reports of Motivation: A Study of Amazon Mechanical Turk in the US and India. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, page 2925–2934, New York, NY, USA, 2012. Association for Computing Machinery.

[7] Hala Assal and Sonia Chiasson. Security in the Software Development Lifecycle. In *Proceedings of the Fourteenth USENIX Conference on Usable Privacy and Security*, SOUPS '18, page 281–296, USA, 2018. USENIX Association.

[8] Hala Assal and Sonia Chiasson. Think Secure from the Beginning: A Survey with Software Developers. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–13, New York, NY, USA, 2019. Association for Computing Machinery.

[9] Sebastian Baltes and Stephan Diehl. Worse Than Spam: Issues In Sampling Software Developers. In *Proceedings of the 10th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*, ESEM '16, New York, NY, USA, 2016. Association for Computing Machinery.

[10] Titus Barik, Justin Smith, Kevin Lubick, Elisabeth Holmes, Jing Feng, Emerson Murphy-Hill, and Chris Parnin. Do Developers Read Compiler Error Messages? In *Proceedings of the 39th International Conference on Software Engineering*, ICSE '17, page 575–585. IEEE Press, 2017.

[11] Hans Christian Benestad, Erik Arisholm, and Dag Ingar Kondrup Sjøberg. How to Recruit Professionals As Subjects in Software Engineering Experiments. Department of Information Systems, Agder University College, August 2005.

[12] Richard E Boyatzis. *Transforming Qualitative Information: Thematic Analysis and Code Development*. Sage Publications, Inc., 1998.

[13] Carolin Brandt and Andy Zaidman. Strategies and Challenges in Recruiting Interview Participants for a Qualitative Evaluation. In *International Workshop on Recruiting Participants for Empirical Software Engineering, co-located with the 44th International Conference on Software Engineering (RoPES - ICSE 2022)*. ROPES, May 2022.

[14] Souti Chattopadhyay, Nicholas Nelson, Yenifer Ramirez Gonzalez, Annel Amelia Leon, Rahul Pandita, and Anita Sarma. Latent Patterns in Activities: A Field Study of How Developers Manage Context. In *Proceedings of the 41st International Conference on Software Engineering*, ICSE '19, page 373–383. IEEE Press, 2019.

[15] Jacob Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

[16] Anastasia Danilova, Alena Naiakshina, Stefan Horstmann, and Matthew Smith. Do You Really Code? Designing and Evaluating Screening Questions for Online Surveys with Programmers. In *Proceedings of the 43rd International Conference on Software Engineering*, ICSE '21, page 537–548. IEEE Press, 2021.

[17] Anastasia Danilova, Alena Naiakshina, and Matthew Smith. One Size Does Not Fit All: A Grounded Theory and Online Survey Study of Developer Preferences for Security Warning Types. In *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering*, ICSE '20, page 136–148, New York, NY, USA, 2020. Association for Computing Machinery.

[18] DeepL GmbH. Deepl, 2022.

[19] F. Ebert, A. Serebrenik, C. Treude, N. Novielli, and F. Castor. On Recruiting Experienced GitHub Contributors for Interviews and Surveys on Prolific. In *The 1st International Workshop on Recruiting Participants for Empirical Software Engineering (RoPES)*, 2022.

[20] Robert Feldt, Thomas Zimmermann, Gunnar R. Bergersen, Davide Falessi, Andreas Jedlitschka, Natalia Juristo, Jürgen Münch, Markku Oivo, Per Runeson, Martin Shepperd, Dag I. Sjøberg, and Burak Turhan. Four Commentaries on the Use of Students and Professionals in Empirical Software Engineering Experiments. *Empirical Softw. Engg.*, 23(6):3801–3820, dec 2018.

[21] Joseph L Fleiss, Bruce Levin, and Myunghee Cho Paik. *Statistical Methods for Rates and Proportions*. John Wiley & Sons, 2013.

[22] Nicolas E. Gold and Jens Krinke. *Ethical Mining: A Case Study on MSR Mining Challenges*, page 265–276. Association for Computing Machinery, New York, NY, USA, 2020.

[23] Marco Gutfleisch, Jan H. Klemmer, Yasemin Acar, Sascha Fahl, and Martina Angela Sasse. Recruiting Software Professionals for Research Studies: Lessons Learned with the Freelancer Platform Upwork. In *International Workshop on Recruiting Participants for Empirical Software Engineering, co-located with the 44th International Conference on Software Engineering (RoPES - ICSE 2022)*. ROPES, May 2022.

[24] Marco Gutfleisch, Jan H. Klemmer, Niklas Busch, Yasemin Acar, M. Angela Sasse, and Sascha Fahl. How Does Usable Security (Not) End Up in Software Products? Results From a Qualitative Interview Study. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 893–910, 2022.

[25] Joseph Hallett, Nikhil Patnaik, Benjamin Shreeve, and Awais Rashid. "Do this! Do that!, and Nothing will Happen" Do Specifications Lead to Securely Stored Passwords? In *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)*, pages 486–498, 2021.

[26] S.E. Hove and B. Anda. Experiences from Conducting Semi-structured Interviews in Empirical Software Engineering Research. In *11th IEEE International Software Metrics Symposium (METRICS'05)*, pages 10 pp.–23, 2005.

[27] Gary Hsieh and Rafał Kocielnik. You Get Who You Pay for: The Impact of Incentives on Participation Bias. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, CSCW '16, page 823–835, New York, NY, USA, 2016. Association for Computing Machinery.

[28] Nicolas Huaman, Alexander Krause, Dominik Wermke, Jan H. Klemmer, Christian Stransky, Yasemin Acar, and Sascha Fahl. If You Can't Get Them to the Lab: Evaluating a Virtual Study Environment with Security Information Workers. In *Eighteenth Symposium on Usable Privacy and Security (SOUPS 2022)*, pages 313–330, Boston, MA, August 2022. USENIX Association.

[29] Harjot Kaur, Sabrina Amft, Daniel Votipka, Yasemin Acar, and Sascha Fahl. Where to Recruit for Security Development Studies: Comparing Six Software Developer Samples. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 4041–4058, Boston, MA, August 2022. USENIX Association.

[30] Mannat Kaur, Michel van Eeten, Marijn Janssen, Kevin Borgolte, and Tobias Fiebig. Human Factors in Security Research: Lessons Learned from 2008-2018, 2021.

[31] Katja Kevic, Braden M. Walters, Timothy R. Shaffer, Bonita Sharif, David C. Shepherd, and Thomas Fritz. Tracing Software Developers' Eyes and Interactions for Change Tasks. In *Proceedings of the 2015 10th Joint Meeting on Foundations of Software Engineering*, ESEC/FSE 2015, page 202–213, New York, NY, USA, 2015. Association for Computing Machinery.

[32] Katharina Krombholz, Wilfried Mayer, Martin Schmiedecker, and Edgar Weippl. "I Have No Idea What i'm Doing": On the Usability of Deploying HTTPS. In *Proceedings of the 26th USENIX Conference on Security Symposium*, SEC'17, page 1339–1356, USA, 2017. USENIX Association.

[33] Thomas D. LaToza, Gina Venolia, and Robert DeLine. Maintaining Mental Models: A Study of Developer Work Habits. In *Proceedings of the 28th International Conference on Software Engineering*, ICSE '06, page 492–501, New York, NY, USA, 2006. Association for Computing Machinery.

[34] Ze Shi Li, Colin Werner, Neil Ernst, and Daniela Damian. GDPR Compliance in the Context of Continuous Integration, 2020.

[35] LinkedIn Corporation. Linkedin, 2022.

[36] Narendar Manohar, Freya MacMillan, Genevieve Z. Steiner, and Amit Arora. *Recruitment of Research Participants*, pages 1–28. Springer Singapore, Singapore, 2018.

[37] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. Reliability and Inter-Rater Reliability in Qualitative Research: Norms and Guidelines for CSCW and HCI Practice. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), nov 2019.

[38] Meta Platforms, Inc. Facebook, 2022.

[39] Abraham H. Mhaidli, Yixin Zou, and Florian Schaub. "We Can't Live Without Them!" App Developers' Adoption of Ad Networks and Their Considerations of Consumer Risks. In *Fifteenth Symposium on Usable Privacy and Security (SOUPS 2019)*, pages 225–244, Santa Clara, CA, August 2019. USENIX Association.

[40] Eyitemi Moju-Igbene, Hanan Abdi, Alan Lu, and Sauvik Das. "How Do You Not Lose Friends?": Synthesizing a Design Space of Social Controls for Securing Shared Digital Resources Via Participatory Design Jams. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 881–898, Boston, MA, August 2022. USENIX Association.

[41] Jefferson Seide Molléri, Kai Petersen, and Emilia Mendes. An Empirically Evaluated Checklist for Surveys in Software Engineering. *Information and Software Technology*, 119:106240, 2020.

[42] Alena Naiakshina, Anastasia Danilova, Eva Gerlitz, and Matthew Smith. On Conducting Security Developer Studies with CS Students: Examining a Password-Storage Study with CS Students, Freelancers, and Company Developers. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–13, New York, NY, USA, 2020. Association for Computing Machinery.

[43] Alena Naiakshina, Anastasia Danilova, Christian Tiefenau, Marco Herzog, Sergej Dechand, and Matthew Smith. Why Do Developers Get Password Storage Wrong? A Qualitative Usability Study. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, CCS '17, page 311–328, New York, NY, USA, 2017. Association for Computing Machinery.

[44] Alena Naiakshina, Anastasia Danilova, Christian Tiefenau, and Matthew Smith. Deception Task Design in Developer Password Studies: Exploring a Student Sample. In *Proceedings of the Fourteenth USENIX Conference on Usable Privacy and Security*, SOUPS '18, page 297–313, USA, 2018. USENIX Association.

[45] Daniela Seabra Oliveira, Tian Lin, Muhammad Sajidur Rahman, Rad Akefirad, Donovan Ellis, Eliany Perez, Rahul Bobhate, Lois A. DeLong, Justin Cappos, and Yuriy Brun. API Blindspots: Why Experienced Developers Write Vulnerable Code. In *Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018)*, pages 315–328, Baltimore, MD, August 2018. USENIX Association.

[46] Open Broadcaster Software. Obs studio, 2022.

[47] Nikhil Patnaik, Joseph Hallett, Mohammad Tahaei, and Awais Rashid. If You Build It, Will They Come? Developer Recruitment for Security Studies. In *International Workshop on Recruiting Participants for Empirical Software Engineering, co-located with the 44th International Conference on Software Engineering (RoPES - ICSE 2022)*. ROPES, May 2022.

[48] A. Rainer and C. Wohlin. Recruiting participants and sampling items of interest in field studies of software engineering. In *The 1st International Workshop on Recruiting Participants for Empirical Software Engineering (RoPES)*, 2022.

[49] Austen Rainer and Claes Wohlin. Recruiting credible participants for field studies in software engineering research. *Information and Software Technology*, 151:107002, 2022.

[50] Brittany Reid, Marcelo d'Amorim, Markus Wagner, and Christoph Treude. NCQ: Code Reuse Support for Node.js Developers. *IEEE Transactions on Software Engineering*, 49(5):3205–3225, 2023.

[51] Brittany Reid, Markus Wagner, Marcelo d'Amorim, and Christoph Treude. Software Engineering User Study Recruitment on Prolific: An Experience Report. In *International Workshop on Recruiting Participants for Empirical Software Engineering, co-located with the 44th International Conference on Software Engineering (RoPES - ICSE 2022)*. ROPES, May 2022.

[52] Norsaremah Salleh, Rashina Hoda, Moon Ting Su, Tanjila Kanij, and John Grundy. Recruitment, Engagement and Feedback in Empirical Software Engineering Studies in Industrial Contexts. *Information and Software Technology*, 98:161–172, 2018.

[53] Dag I. K. Sjøberg, Bente Anda, Erik Arisholm, Tore Dybå, Magne Jürgensen, Amela Karahasanovic, Espen F. Koren, and Marek Vokác. Conducting Realistic Experiments in Software Engineering. In *Proceedings of the*

2002 International Symposium on Empirical Software Engineering, ISESE '02, page 17, USA, 2002. IEEE Computer Society.

[54] Edward Smith, Robert Loftin, Emerson Murphy-Hill, Christian Bird, and Thomas Zimmermann. Improving Developer Participation Rates in Surveys. In *2013 6th International Workshop on Cooperative and Human Aspects of Software Engineering (CHASE)*, pages 89–92, 2013.

[55] Leonardo Sousa, Anderson Oliveira, Willian Oizumi, Simone Barbosa, Alessandro Garcia, Jaejoon Lee, Marcos Kalinowski, Rafael de Mello, Baldoino Fonseca, Roberto Oliveira, Carlos Lucena, and Rodrigo Paes. Identifying Design Problems in the Source Code: A Grounded Theory. In *Proceedings of the 40th International Conference on Software Engineering*, ICSE '18, page 921–931, New York, NY, USA, 2018. Association for Computing Machinery.

[56] Kathryn T. Stolee and Sebastian Elbaum. Exploring the Use of Crowdsourcing to Support Empirical Studies in Software Engineering. In *Proceedings of the 2010 ACM-IEEE International Symposium on Empirical Software Engineering and Measurement*, ESEM '10, New York, NY, USA, 2010. Association for Computing Machinery.

[57] Christian Stransky, Yasemin Acar, Duc Cuong Nguyen, Dominik Wermke, Elissa M. Redmiles, Doowon Kim, Michael Backes, Simson Garfinkel, Michelle L. Mazurek, and Sascha Fahl. Lessons Learned from Using an Online Platform to Conduct Large-Scale, Online Controlled Security Experiments with Software Developers. In *Proceedings of the 10th USENIX Conference on Cyber Security Experimentation and Test*, CSET'17, page 6, USA, 2017. USENIX Association.

[58] Mohammad Tahaei, Alisa Frik, and Kami Vaniea. Privacy Champions in Software Teams: Understanding Their Motivations, Strategies, and Challenges. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery.

[59] Mohammad Tahaei and Kami Vaniea. Lessons Learned From Recruiting Participants With Programming Skills for Empirical Privacy and Security Studies. In *International Workshop on Recruiting Participants for Empirical Software Engineering, co-located with the 44th International Conference on Software Engineering (RoPES - ICSE 2022)*. ROPES, May 2022.

[60] Mohammad Tahaei and Kami Vaniea. Recruiting Participants With Programming Skills: A Comparison of Four Crowdsourcing Platforms and a CS Student Mailing List. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery.

[61] Christian Tiefenau, Emanuel von Zezschwitz, Maximilian Häring, Katharina Krombholz, and Matthew Smith. A Usability Evaluation of Let's Encrypt and Certbot: Usable Security Done Right. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, CCS '19, page 1971–1988, New York, NY, USA, 2019. Association for Computing Machinery.

[62] Twitter, Inc. Twitter, 2022.

[63] Sira Vegas, Óscar Dieste, and Natalia Juristo. Difficulties in Running Experiments in the Software Industry: Experiences from the Trenches. In *2015 IEEE/ACM 3rd International Workshop on Conducting Empirical Studies in Industry*, pages 3–9, 2015.

[64] VERBI Software. Maxqda, 2022.

[65] XING SE. Xing, 2022.

[66] Khaled Yakdan, Sergej Dechand, Elmar Gerhards-Padilla, and Matthew Smith. Helping Johnny to Analyze Malware: A Usability-Optimized Decompiler and Malware Analysis User Study. In *2016 IEEE Symposium on Security and Privacy (SP)*, pages 158–177, 2016.

[67] Zoom Video Communications, Inc. Zoom, 2022.

A Pre-Survey

A.1 Consent

I agree to take part in this survey
○ I agree / I do not agree

A.2 Questions

Q1: How old are you?

Q2: What best describes your current primary occupation?
○ Employee / Freelancer / Researcher / Apprentice / Bachelor student / Master student / Other (please specify:)

Q3: Are you employed by a company and work more than 19 hours a week? (Yes / No)

Q4: In which area do you currently work?
○ Consulting / Software Development / Testing / Other (please specify:)

Q5: Is software development part of your job?
○ Yes / No / Other (please specify:)

Q5: Which email address may we use to contact you regarding your participation in our study?

B Post-Survey

B.1 Consent

I agree to take part in this survey
○ I agree / I do not agree

B.2 Questions

Q1: Please enter your pseudonym:

Q2: How old are you?

Q3: What is your gender?
○ Male / Female / Diverse / I prefer not to answer / I prefer to describe myself:

Q4: What is your nationality?

Q5: How long in total have you been employed as a software developer in a German company or organization? (In years)

Q6: How many hours per week do you spent developing software?

Q7: What is your job title?

Q8: What kind of software are you developing? (Multiple selections possible)
○ Web applications / Mobile applications / Desktop applications / Embedded software development / Enterprise applications / Other (please specify:)

Q9: How many years of experience with software development do you have in general?

Q10: How many employees has your company in total?
○ 1-9 / 10-249 / 250-499 / 500-999 / 1000 or more

Q11: Which business sector does your company belong to?
○ Game development / Building network and communication / Web development / Development of middleware, system components, libraries and frameworks / Other (please specify:)

Q12: Has your company a focus on security? (Yes / No)

Q13: Are you taking on security-related tasks in your field of work? (Yes / No)

Q14: Do you want to be contacted by our research group to be informed about the study results? (Yes / No)

Q15: Do you want to be invited by our research group for further studies? (Yes / No)

Q16: If you would like to be contacted by our research group, please provide a valid e-mail address. This e-mail address will be stored separately from the study data.

C Interview Guideline

Introduction

- Have you participated in a scientific study in the past?
  - *If yes:* Was this a developer study?
- In how many studies have you taken part in the past?
- In what kind of studies have you taken part in the past?
  - *Researcher note: Explicitly ask for programming tasks*
- Have you ever had a negative experience?

Influence - Study Topic

- Is the study topic something you take into account when deciding whether to participate in a study? Is the topic of the study important to you?
  - *If yes:* What topics are you interested in?
  - *Computer Science/Software Development/IT Security/private interests as a study topic*
- Would you also participate in studies on topics that are not interesting to you?
  - *If No:* Could you be convinced to participate in a study on a topic not interesting to you? *If Yes:* How?
  - *If job-related studies are not attractive:* How could you be motivated for this kind of study?

Influence - Trust

- During a study, information about your person is collected. The researchers conducting the study must adhere to rules (e.g., the anonymity of participants). Participants are informed about this in a consent form.

- To what extent do you trust a university to comply with informed consent?
  - *Researcher note, Address the following aspects and ask how important they are:* Respect for privacy and data protection, fair compensation, competence
- Does your opinion change if the conducting organization is from a private sector? *(Researcher note: Address the factors mentioned above again)*
- What do you think about organizations such as the Fraunhofer-Gesellschaft?
  - *Researcher note, alternatives to Fraunhofer: Max Planck Gesellschaft, German Research Foundation (DFG)*
- Does the type of organization impact the compensation you expect?
  - *If yes:* To what extent does the type of organization influence your expectation?

Influence - Researcher Background

- What influence does the background of the conducting researchers have on your willingness to participate in a study?
  - *Researcher note, examples: Computer scientist, psychologist, social scientist*
- Does the background have to match the topic?
- Would you prefer/avoid one of the mentioned backgrounds?
  - *If yes:* Why?

Influence - Lab, Field and Online Studies

- *(Researcher note: Shortly explain the difference between lab, field, and online studies.)* Do you understand what I mean by this?
- What do you (not) like about online studies?
- What do you (not) like about lab studies?
- What do you (not) like about field studies
- Would this form of study be conceivable in your company?
- Which type of study do you prefer?
  - Why do you prefer this type of study?
  - What are the arguments against the other types of studies?
- Does the type of study influence the compensation you expect?
  - *Researcher note: Type of study in terms of the difference between lab, field, and online studies*
  - *If yes:* To what extent does the type of study influence your expectation?

Influence - Study Task

- During a developer study, there are usually three possible types of tasks. A survey, an interview, and practical tasks such as coding, writing a code review, or a protocol. Does the type of task in a study influence your willingness to participate in a study?
- Which tasks do you like to work on the most? Why?
- Are there tasks you avoid? Why?
- *If not already addressed:* How do you perceive the difficulty of these types of tasks?
  - What makes a task difficult for you?
  - What makes a task easy for you?
- Does the type of study task influence the compensation you expect for the same duration? *If yes:* To what extent does the type of study task influence your expectation, given the same duration?

Influence - Duration

- The duration of a study can vary depending on the research question. Some studies consist of a short questionnaire, and other studies can take several days.
- Does the duration of a study influence your willingness to participate? Here we assume you receive the same hourly compensation regardless of the study duration.
  - Is duration a criterion by which you exclude a study?
  - *Researcher note: If not mentioned, address the following aspects:* Conflict with working time, conflict with free-time, fatigue
- Is there a maximum duration beyond which you would no longer be willing to participate in a study?
  - Do other study factors influence this maximum duration? Why?
- Under what conditions would you be more likely to participate in long studies?

  - *Researcher note, for example:* better compensation, less organizational effort, provision of catering, payment for travel.

- Would you be willing to participate in studies lasting several days, provided that your maximum duration is not exceeded on a individual days?
  - *If no:* Why?
- Is there a maximum number of days or a maximum time period?
- Does the duration of a study influence the compensation you expect?
  - *If yes:* To what extent does the duration of a study influence your expectation? Why?

Influence - Data Collection

- During a study, information about you is collected, such as your voice or image. Depending on the study, written code or biometric characteristics may also be recorded.
- Are you concerned about the data mentioned if it is collected from you?
  - *If yes:* Which specific data are you concerned about?
- Does the way it is recorded have an impact on your willingness to participate?
  - *If yes:* Is there any data, if recorded, that would stop you from participating in a study?
- Would you avoid studies that record data you are concerned about?
- Do you expect better compensation if the study records data you are concerned about?

Influence - Compensation

- How would you like to be compensated for the participation in developer studies?
  - *Researcher Note, for example:* Workshop participation, software license, conference tickets
- How do you feel about non-monetary compensation?
- Would you accept less compensation if the study would meet your preferences for study factors?
- Would you also be prepared to forego compensation completely?
  - *If yes:* Which study factors are particularly important to you in this regard?

Recruitment

- There are two methods for recruiting study participants. In active recruitment, potential participants are contacted directly by the researcher. In passive recruitment, the study is advertised in various ways, and a potential participant contacts the researcher independently.
  - What advantages (disadvantages) do you expect from active recruitment?
  - What advantages (disadvantages) do you expect from passive recruitment?
  - In which way would you prefer to be recruited?
- Where would you prefer to be recruited? Why?
  - *Researcher note, for example:* At work, on-site events, online events, by email, online communities such as Facebook group/Reddit/Xing/LinkedIn, flyers (by post/by hand), posters, employer
  - How should the first contact be made?
  - *Scenario:* An unknown university sends an email to you. How do you feel about that? How do you react?
  - Would an explanation on how the contact information was collected defuse the situation?
- Would you be willing to register to a website for future study invitations?
- Who should manage this platform?

Motivation

- What is the main reason that motivates you to participate in a developer study?
  - *Researcher note: Ask for other reasons besides the main reason*
  - *Researcher note: Address this and previous studies:* What were the reasons for participating in these/this studies?
- Are there any study factors that influence your willingness to participate that have not been mentioned yet? *Researcher note: If the participant has already participated in several studies, ask what researchers could do better*

D Codebook

Table 3: Codebook used during the analysis of the transcribed interviews using thematic analysis

Theme	Code	Definition
Recruitment	**Channel**	**Channel:** All statements made regarding recruitment channel (where participants want to be recruited) and their influence on a participant's willingness to participate
	Active	**Active:** All statements made about active recruitment and its influence on a participant's willingness to participate
	Passive	**Passive:** All statements made about passive recruitment and its influence on a participant's willingness to participate
Compensation	**Compensation**	All statements about compensation as the reward for participation, which is promised before participation. Includes non-monetary compensation such as software, hardware, vouchers, vacation trips or education offers
Length	**Length**	Direct and indirect statements about the influence of the study duration on a participant's willingness to participate and the compensation they expect. This includes statements about time flexibility
Study Task	**Study Task**	All statements about the study task (survey, interview, practical tasks) and their influence on a participant's willingness to participate and the compensation they expect
Study Type	**Study Type**	All statements about the study type (online, lab, field) and their influence on a participant's willingness to participate and the compensation they expect
Trust in Organization	**Trust in Organization**	**Trust in Organization:** All statements about the conducting organization (university, company, hybrid institutions) and their influence on a participant's willingness to participate and the compensation they expect. This includes trust in the intention of the conducting organization as well as trust in the competence of the conducting organization
	Data Collection	**Data Collection:** Opinions, fears, apprehensions, problems and wishes in the context of data collection. All statements on the influence of data collection on a participant's willingness to participate and the compensation they expect
Motivation for Participation	**Motivation for Participation**	All statements made explicitly about the influence on a participant's willingness to participate in a study that uses the keywords *participate* or *participation*
Influence Study Topic	**Influence Study Topic**	**Influence Study Topic:** All statements made about specific study topics, such as computer science, IT-security or subtopics of these and their influence on a participant's willingness to participate and the compensation they expect
	Researcher Background	**Researcher Background:** All statements about a researcher's background and their influence on a participant's willingness to participate and the compensation they expect. This includes trust in the competence of the individual researcher background

SoK: I Have the (Developer) Power! Sample Size Estimation for Fisher's Exact, Chi-Squared, McNemar's, Wilcoxon Rank-Sum, Wilcoxon Signed-Rank and t-tests in Developer-Centered Usable Security

Anna-Marie Ortloff
University of Bonn

Christian Tiefenau
University of Bonn

Matthew Smith
University of Bonn, Fraunhofer FKIE

Abstract

A priori power analysis would be very beneficial for researchers in the field of developer-centered usable security since recruiting developers for studies is challenging. Power analysis allows researchers to know how many participants they need to test their null hypotheses. However, most studies in this field do not report having conducted power analysis. We conducted a meta-analysis of 54 top-tier developer study papers and found that many are indeed underpowered even to detect large effects. To aid researchers in conducting a priori power analysis in this challenging field, we conducted a systematization of knowledge to extract and condense the needed information. We extracted information from 467 tests and 413 variables and developed a data structure to systematically represent information about hypothesis tests, involved variables, and study methodology. We then systematized the information for tests with categorical independent variables with two groups, i.e., Fisher's exact, chi-squared, McNemar's, Wilcoxon rank-sum, Wilcoxon signed-rank, and paired and independent t-tests to aid researchers with power analysis for these tests. Additionally, we present overview information on the field of developer-centered usable security and list recommendations for suitable reporting practices to make statistical information for power analysis and interpretation more accessible for researchers.

1 Introduction

A priori power analysis can be used to calculate the necessary sample size for a study to detect an effect with a given probability, a given significance criterion, and a defined effect size [22]. The probability is often set to 80% and the significance criterion to 5% by convention and the effect size needs to be chosen by the researcher based on their research goals [22, 30]. Using a priori power analysis, researchers can avoid running underpowered studies and missing effects that are actually present in the population and thus wasting resources or, worse yet, potentially publishing results that are misinterpreted as stating that there is no effect. It also prevents researchers from using more resources than necessary by running overpowered studies [35]. This is especially problematic for the sub-field of developer centered usable security (DCUS) [48] since developers are often both hard and expensive to recruit. Running underpowered studies in this field is especially undesirable due to the large amount of effort coupled with low chances of finding the desired effects even if they are there. Despite this, power analysis is not common in the field of DCUS.

In the 54 DCUS papers we analyzed for this SoK, only 9,3% contained any form of power analysis. When the power of studies has been assessed in other fields in meta-analyses, these have frequently shown low power, such as in a review of ACM transactions [12] or in psychology [99]. The lack of a priori power analysis is one likely reason for this. Our analysis raises similar concerns of underpowered studies in DCUS. A potential reason for so few studies being planned with power analysis is that performing a power analysis is non-trivial and researchers must know, estimate or guess key population values such as standard deviations or proportions to calculate effect sizes or estimate the effect sizes themselves. This is especially tricky in the fields of Usable Security and Privacy (USP) and DCUS, due to both the heterogeneity of the populations studied, as well as the heterogeneity of variables being measured and the many non-standardized measurement instruments. On top of that, many tests are not published with enough statistical details to use them for estimating power for similar future studies [50].

In this paper, we present a systematization of knowledge in the field of DCUS with the goal to aid researchers with

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.
USENIX Symposium on Usable Privacy and Security (SOUPS) 2023. August 6–8, 2023, Anaheim, CA, USA

power calculations for some common statistical tests used in the field: tests with a single categorical independent variable with two groups, with either a categorical dependent variable: Fisher's exact test, chi-squared test, and McNemar's test, or with a continuous dependent variable: independent and paired t-test, as well as Wilcoxon rank-sum test and Wilcoxon signed-rank test. We collected DCUS papers from the following major conferences published between 2010 and 2021: the Symposium on Usable Privacy and Security (SOUPS), USENIX Security, the IEEE Symposium on Security and Privacy (S&P), the ACM Conference on Computer and Communications Security (CCS), the IEEE/ACM International Conference on Software Engineering (ICSE) and the security and privacy sessions from the ACM Conference on Human Factors in Computing Systems (CHI). We excluded any paper that did not contain an actual user study. This left us with a set of 54 papers. From these, we manually extracted all relevant data, including information about 467 statistical hypothesis tests involving 413 different variables. We developed a data structure to make this information accessible to researchers for specific power analyses. We further systematized the data from the papers by categorizing the involved variables into 13 different groups to make it easier to find proxies for power calculations if a direct match is not available. We also use our categories to calculate average statistics which can help researchers sanity check their estimates. The database including all the data will be made available to the community. Based on our findings, we make recommendations on how to conduct power analysis for developer studies.

2 Background & Related Work

In the following, we give a very brief overview of the research domain of developer centered usable security before discussing the background of power analysis. We also examine the application of power analysis and the practice of reporting effect sizes and conducting meta-analyses since these are closely related to power analysis.

When examining the use of statistical techniques, such as power analysis, meta-analysis, or reporting of effect sizes, in the following, we try to summarize the state of the practice as close to DCUS as possible. However, since this is a relatively new field [48] there is not yet much work providing an overview of the use of such techniques. Instead, we examine related domains, such as USP and Human Computer Interaction (HCI) in general, or when this is not possible, psychology. While these fields are by no means equal, we posit that methods and constraints are at least comparable in that user studies measuring latent variables, and not only directly observable variables, are common in all of the mentioned fields.

2.1 Developer Centered Usable Security

DCUS is a subfield of USP, but with a focus on the challenges and needs of expert users, such as software developers or administrators, instead of end users, who are at the center of typical USP-studies [3, 48]. DCUS extends USP's notion that security mechanisms should be designed with users in mind, to developers, which are themselves users, e.g., of cryptography APIs [1, 48, 73], programming languages [90] and other security tools [e.g. 9, 61, 88]. Tahaei & Vaniea provide an overview of topics addressed and methods used in DCUS [103]. While developers have been the main focus, we also include other expert users such as administrators in this field (e.g., [106].)

2.2 Theoretical Background of Power Analysis

Power analysis as a concept is situated within the null hypothesis significance testing (NHST) paradigm of statistical analysis. Four parameters are relevant to power analysis: Power, i.e., the probability of the test correctly rejecting the null hypothesis, the significance criterion α, the reliability of the sample results, and the effect size [22, 35]. The largest and invariably present influencing factor on reliability is sample size [22] - larger samples produce more consistent and reliable estimates than smaller ones. Consequently, the sample size is often used as a stand-in for reliability in power analysis.

These four parameters are interdependent, such that when three of them are available, it is possible to calculate the fourth. These calculations are referred to as power analysis. In general, there are four different kinds of power analysis, each used to determine one of the parameters from the other three [22]. The focus of this work is on so-called a priori, or prospective power analysis, which is used to calculate the necessary sample size to detect an effect of a desired size with a chosen power and significance criterion, before actually conducting the study. For a more detailed introduction to the four parameters, see Appendix C or Ellis (2010) [35].

To be able to conduct an a priori power analysis, researchers need to know or guess either a standardized effect size they expect to detect or related statistics that can be used to calculate such an effect size. These effect sizes [35] and the resulting power analysis procedures [38] vary depending on the test used. The standardized effect sizes needed for the tests at the focus of our work are ϕ for the chi-squared test, the odds ratio for McNemar's test, Cohen's d for independent samples, and Cohen's d_z for paired samples mean-comparison tests. While effect sizes can be converted between each other, what is generally seen as small, medium, or large differs between the effect size types [22], and this makes the process of guessing more difficult since not every researcher is familiar with the same types of effect sizes. Some procedures require different information, such the Fisher's exact test, where success prob-

abilities for both groups are needed instead of a standardized effect size, or McNemar's test, which requires the proportion of cases, where changes occur in subjects' responses (proportion of discordant pairs) in addition to the odds ratio effect size. Alternatively, researchers can also guess unstandardized effect sizes, such as group means for both groups, i.e., the difference between these means. This can be more intuitive, especially when taking the approach of aiming to detect the smallest practically relevant effect sizes [35]. However, additional statistics are needed to calculate the necessary standardized effect size from these values. For the tests we focus on in this work, these are the standard deviations for the two groups for independent and paired t-tests, Wilcoxon rank-sum tests, and Wilcoxon signed-rank tests. Additionally, the correlation between the two groups is necessary for within-subjects tests with continuous dependent variables. The main aim of this paper is to help and guide researchers to base these guesses on previous work or at least enable sanity checks on estimated values.

A priori power analysis is important, as both under- and overpowered studies are detrimental to the furthering of knowledge. Conducting an underpowered study means that failing to reject the null hypothesis is likely [35]. Since non-significant results are less likely to be published [7, 96], the effort in planning and conducting the underpowered study may be wasted. On the other hand, overpowered studies are wasteful, too [35]. Highly powered tests can detect very small effects so that in extreme cases, it is possible to find a highly statistically significant, albeit very small actual difference, which may be irrelevant in practice. Less power and fewer resources would have been sufficient to detect a practically relevant effect [35]. In addition to waste of resources, only collecting data from as many participants as necessary minimizes the amount of data collected, with positive effects on participants' privacy. Additionally, both underpowered and overpowered studies may be interpreted incorrectly when focusing on p-values. Underpowered non-significant results may be dismissed as irrelevant, even in the case of a large effect, while overpowered significant results representing trivial effects can be posited as important due to the statistical significance [35]. In DCUS, it is especially important to be mindful when recruiting since developers as specialists are usually time-constrained, and payment is often much higher than in end-user studies.

2.3 Application of Power Analysis

The tools to conduct power analysis have evolved from power and sample size tables [22] to online calculators[1], designated computer programs, like G*Power [38] and multiple implementations in programming languages commonly used for statistical analysis, like the `pwr` package, among others in R [91] or the `statsmodels` library in Python [98].

[1] e.g., powerandsamplesize.com or jakewestfall.org/power

Nevertheless, historically, power analysis has often not been applied [14, 22]. Unfortunately, in many fields, this is still the case according to more recent reviews, such as in psychology, where only 5% of 183 reviewed publications mentioned power analysis [113]. In other fields, often in the medical domain, power analysis is more prevalent, e.g., 43% of studies in a review of obesity interventions in schools [55] and over 60% of reviewed publications in NEJM and Lancet reported prospective power analyses [109]. A possible reason is adherence to submission guidelines [108], which we recommend updating for the field of USP. If there is enough prior information, power analysis is recommendable for grant applications to ensure sufficient funds are planned for recruitment.

In HCI, power analysis is also frequently not applied. E.g., in 2018, only five of 519 experimental papers at CHI used prospective power analysis [34]. In interviews evaluating a prototype of a program to facilitate power analyses, some researchers were explicitly skeptical of power analysis as a research tool [34]. In a more cursory evaluation of terminology used in the CHI proceedings of 2017 - 2019, rather than manual inspection of publications, only between 1.5% and 2.7% of the papers containing the term "experiment" also contained the term "power analysis" [33]. In our own analysis of USP publications at SOUPS and CHI in 2021 and 2022, we found that only 5.4% of SOUPS papers used power analysis in some form and 8.3% of USP CHI papers did so. Over these two years, only ten of 146 (6.8%) USP papers at CHI and SOUPS used power analysis in some way. Two of those papers conducted post hoc power analysis, which is controversial since power and p-values are directly related, and if a result is not significant, i.e., the p-value is high, then power was too low to detect an effect of the size present in the sample. So little is gained by the post hoc power calculation [45, 56].

Consequently, power to detect small effects in published studies is frequently low [35]. Literature shows this to be the case for such diverse research areas as management information systems, including ACM transactions [12], psychology [93, 99], and health professions education [25]. While not a formal review of power, Cockburn et al. examine empirical computer science literature and find evidence of the same practices that contributed to the replication crisis in other domains [21]. At the largest HCI conference, CHI, quantitative studies may even be underpowered to detect large effects [20].

When power analysis is not used to determine appropriate sample size, alternative approaches include recruiting the maximal number of participants possible, based on population, time, and monetary constraints [35], following prior practice and experience in the domain of interest [20, 35] or using rules of thumb, such as 10 or 15 cases of data per predictor [39], $50 + 8 \times k$, where k is the number of predictors in regression for testing the overall model [49], or two subjects per variable to estimate the coefficients in linear regression [8].

However, none of these methods guarantee that studies will have sufficient power [35].

To increase power, especially in fields like DCUS, where recruiting is challenging, researchers should also consider adapting their research design to increase the reliability of measurements and reduce random errors, e.g., by conducting within-subjects research [35]. Another possibility is conducting sequential analyses, whereby a study can be stopped at planned intervals if a large enough effect can be detected at this time, which on average reduces the necessary number of participants [67]. There is some discussion about whether stopping studies early like this introduces an additional bias towards larger effects [13, 72] or not [42, 97].

2.4 Effect sizes and Meta analysis

Reporting of effect sizes and conducting meta-analyses are topics closely related to power and power analysis. Researchers need to determine an expected effect size or a minimum relevant effect size to conduct power analyses. Meta-analyses, in contrast to (systematic) literature reviews, which provide a narrative summary of a research domain, make use of effect sizes to combine findings from different studies [35]. In conducting re-analyses with larger amounts of data, they achieve a higher power to detect effects [35].

A general lack of detail in statistical reporting was admonished in early HCI meta-analyses [79], and lack of effect size reporting is a problem, e.g., in studies investigating software engineering [62]. Groß's analysis of statistical reporting in USP showed that half of the 114 analyzed user studies from 2006–2016 reported incomplete results [50]. This makes both prospective power analysis and meta-analyses more difficult [50, 62]. In general, there are few meta-analyses in HCI [64]. As a domain using diverse analysis methods and tools derived, e.g., from computer science, design science, or psychology, there are not necessarily unified reporting standards within HCI, which makes meta-analyses difficult [104]. When reporting is not sufficient, a workaround is to ask authors for the raw data, but this comes with the drawback that not all authors will respond, e.g., Hornbaek et al. had a response rate of 48% [57]. Nevertheless, there are examples of meta-analyses in HCI, e.g., about human-robot interaction [36], usability measures [57], and typing experiments [81]. While there are certainly literature reviews in USP, e.g., [15, 32] and also in DCUS [103] we did not find a meta-analysis in this domain.

As part of our systematization of knowledge, we conducted a meta-analysis concerning statistical power in the field of DCUS. However, due to incomplete reporting and test-specific limitations, we could only do this meta-analysis for 140 tests in 20 studies.

3 Literature Collection

In the following, we describe the creation of our literature corpus for DCUS. As a catch-all, we will refer to this sort of literature as a developer paper in the remainder of this work. We define a developer paper as literature including a user study in some form, in which the participants are software developers, software testers, administrators, other people responsible for planning, developing, testing, or managing software, or proxies for such people. An example of proxies would be computer science students, which are commonly used as a stand-in, e.g., for software developers [77]. We focus on the domain of usable security and privacy, which means that the studies should be focused on privacy or security problems and technology. We exclude any papers which do not include a study with actual users.

We started collection of literature in early 2021 and collected developer papers from four major conferences about security and privacy, which were published between 2010 and 2020: SOUPS, USENIX Security, S&P, CCS, and additionally ICSE and the USP tracks of ACM SIGCHI. Abstracts were used to determine whether a paper fits our definition of a developer paper, and in case of uncertainty, the method section of the paper was additionally used to clarify. We updated our literature basis in March 2022. Our final sample consists of 54 papers. Of those, 20 were published at SOUPS, 11 at CCS, 8 at ICSE, 7 at USENIX Security, 5 at S&P and 3 at CHI. The list can be found in Appendix A.

4 Systematizing Study and Statistical Test Data

Our goal was to collect and systematize information on studies and statistical tests from the domain of DCUS, focusing on what is necessary to conduct power analyses for user studies in this domain. We also wanted to add general information on the data collection process and the types of participants, since this might also be relevant when planning a new study. To aid researchers in planning new studies we created a data structure of our systematization and will offer this to the research community as a database. The database can be queried via a companion website for the relevant information to conduct power analysis [2]. An excerpt of two entries can be found in table 2. Further entries are on the companion website. The entries are categorized to help researchers query the database and find similar studies, on which they can base their power calculation. For those cases where no directly similar previous study exists, we have created aggregated data based on our systematization that researchers can use as rough guides.

[2] https://powerdb.info

4.1 Systematization Process

Based on a sample of the literature we had collected, we first analyzed papers from a methodological point of view and collected information on data collection, data analysis, as well as meta information that served to clearly identify and reference the paper. We identified similarities in the type of collected data and iteratively developed a data structure to represent this information.

In addition to the papers themselves, two sources further informed our structure: We made sure to represent information necessary to conduct power analysis, based on the G*Power-software [38], since G*Power is a commonly used and very powerful tool for power analysis.

To help guide our work, we created a set of hypothetical developer studies, for which we would want to run a priori power calculations, e.g., *Do Freelancers recruited from Freelancer.com and Upwork differ in their self-assessment of their reverse engineering skill and in their performance while completing a short reverse engineering task?*. Our aim was to have a mix of hypothetical studies which were closely related to previous work as well as some that had no relations. This was done to a) ensure that it would be easy to find very specific data from closely related work as well as to b) ensure that our categories were useful to guide researchers in uncharted territory as best possible. Based on the case studies, we added features to categorize variables. Finally, after laying the theoretical foundation, we implemented a database and started to enter information from the collected literature.

4.2 Data Structure

In the following we describe how we systematized the data we collected, providing a general overview of the structure, as well as details on those topics specifically relevant to power analysis and finding the right data.

4.2.1 Overview

A general overview of the data structure can be seen in the entity relationship diagram (ERD) in the appendix B.

For each paper in our set, we first collected meta information about it. Each paper can have multiple studies assigned to it. A *study* is a self-contained unit of a combination of data collection and analysis, which is often presented in a separate section in a paper. We separate *data collection descriptions* from the *participant samples* involved. To support filtering and generalization, *Participant sample types*, instances of which could be "student", "security expert", "freelance developer", and *data collection methods*, where instances are, e.g., "interview", "survey" or "experiment/task-based evaluation", are represented as separate entities. These are more easily reused across multiple studies and multiple papers. The results of qualitative analyses cannot directly be used for power analysis or to support meta-analysis since effect sizes are not calculated. However, because insights from qualitative analysis often help inform further quantitative work, we also collected some information on the *qualitative analysis methods* used. Reporting of NHST-type analysis methods all share certain properties, which we collected for all *quantitative analysis methods*, i.e., the name of the hypothesis test, the p-values, and the dependent and independent variables. For more information on the representation of variables, see Section 4.2.2. We categorized the different hypothesis tests used in our sample according to the number and type of dependent and independent variables, the study design, and, in the case of categorical variables, the number of levels in the variable, see Table 1. In the following, we focus on those hypothesis tests with one categorical independent variable with two levels. We collected additional test-specific data for these tests, see Section 4.2.3.

4.2.2 Representation of Variables

For *variables*, we noted the name and a description of how the variable was measured based on the publication and collected information on several additional facets of the variables. For example, the *variable type* attribute represents whether a variable is continuous or categorical. *Variable levels*, i.e., groups, are then provided for categorical variables.

Finally, a *Variable* can be tagged with one or multiple *Variable categories*. These are broad categories of variables, which frequently occur in studies in DCUS, and which serve to ease the search for a specific variable or test and were used to create generic guides for researchers when no specific prior work exists. We generated these categories by open coding all the variables in eight of the papers from our sample, which we chose to cover a broad range of topics and both descriptive and inferential work. One researcher did all the coding alone, and the generated categories and the corresponding variables were frequently discussed together in an iterative process with a second researcher and modified when necessary. Since we were assigning fixed categories to clear units of data, and a second researcher was involved in the generation of the variable categories, we consider this data simple to code and independent recoding to be unnecessary [70, 84]. The eleven categories which emerged from this process are:

usability Measures relating to overall usability, i.e., incorporating effectiveness, efficiency, and user satisfaction according to ISO 9241-11 [59].

security Measures relating to IT security of produced software / artifacts.

functionality Measures relating to the functionality of produced software / artifacts.

participant judgment Measures relating to participants' choices or judgment of something.

Type of IV	# IV	# IV Levels	Type of DV	# DV Levels	Study design	Hypothesis test
Categorical	One	Two	Categorical	Two	Between	Fisher's Exact Test, Chi-Squared Test
Categorical	One	Two	Categorical	Two	Within	McNemar's Test
Categorical	One	Two	Continuous	-	Between	Independent t-test, Wilcoxon Ranksum Test
Categorical	One	Two	Continuous	-	Within	Paired t-test, Wilcoxon Signed Rank Test
Categorical	One +	Two +	Continuous	-	Between	ANOVA, Kruskal-Wallis Test (1 IV)
Categorical	One +	Two +	Continuous	-	Within	Repeated Measures ANOVA, Friedman's ANOVA (1 IV)
Continuous	One	-	Continuous	-	Between	Pearson correlation, Kendall's tau, Spearman's correlation, Polychoric correlation
Any	One +	Any	Categorical	Two	Between	Logistic Regression
Any	One +	Any	Continuous	-	Between	Linear Regression, Poisson Regresion (DV: counts)
Any	One +	Any	Any	Any	Any	Generalized Linear Mixed Model
fixed value	-	-	Categorical	Two	-	Binomial Test
fixed value	-	-	Continuous	-	-	Z-Test

Table 1: Hypothesis tests appearing in DCUS papers. In this paper we focus on tests with single categorical IV and single DV (highlighted in pink) for our assessment of reporting and effect sizes. Remaining tests have a green background. All tests in our sample had a single dependent variable. Three tests did not fit into this categorization: Bernoulli trial, factor analysis and k-means cluster analysis. DV = dependent variable, IV = independent variable

experience Participants' level of experience in something, e.g., programming.

behavior Measures of participants' behavior. Can be either self-reported or objectively measured.

system type Usually an assigned condition in a study, the system / software / prototype, which participants work with or test.

participant type Group to which a participant belongs, e.g., students, freelancing developers.

participant characteristic Any other participant trait, such as the type of company a participant works for, a participant's focus on security, etc.

task related variable Variables related to the tasks in a study, e.g., which task the participants worked on or task order.

study related variable Measures relating to administrative aspects of the study, e.g., drop-outs, prompting, or additional communication with participants, e.g., via email or a support system.

During the data input process, we added one additional category:

artifact-related variable Variables related to artifacts participants produced during or prior to the study, e.g., characteristics of these or types of mistakes encountered in submitted code or other artifacts.

4.2.3 Relevant Test-specific Information for Power Analysis

We specifically focused on collecting information that would be needed to conduct a priori power analysis, as well as additional values, which are typically reported with a hypothesis test according to APA style [5]. For some tests, not all data could be contained in a single entity, e.g., ANOVA, linear regression, and logistic regression. For these, we created multiple related entities in our database, but the meta-analysis will be carried out in future work since very few of these tests were reported with enough detail for a robust analysis at this point.

4.3 Data Input and Checking

Two assistant researchers were hired to aid the main author in entering data into the database. Both had attended at least one university course on statistical hypothesis testing and empirical methods. The main author also has experience teaching empirical methods and statistics.

The main author trained the other two researchers regarding data extraction from the papers and how the data should be entered into the database. The data entry tool provided a checkbox that could be used to mark an entry when feeling unsure, as well as a text field where the issue could be noted. The two assistant researchers received feedback regarding their data entry at set intervals. At the end of the data input process, the main author went over all entered data again to check for any missing data or inconsistencies. Uncertain cases were discussed and resolved together with the co-authors.

5 Meta-Analysis

In the following, we analyze and further systematize the data we gathered. For our meta-analysis of the current state of research in DCUS, we focus on information related to power analysis, e.g., we analyze effect sizes in this field and investigate whether reporting is sufficiently detailed to enable power analysis using the data. We analyzed the data from our database using the Python packages numpy, pandas, and matplotlib and the R tidyverse [91, 119].

We analyzed a total of 54 developer papers, which encompassed 64 individual studies, of which 24 were quantitative, 24 were qualitative, and 16 used both quantitative and qualitative analysis methods. On average, 105 (9 - 330, median=65, SD=99.5) participants took part in quantitative and 14.8 (1 - 49, median=12, SD=12.3) in qualitative studies, and for mixed methods studies on average, 103 (6 - 400, median=44, SD=101.6) participants took part. This is similar to Caine's analysis of papers at CHI 2014, where the sample size was also smaller for qualitative than quantitative work [20].

5.1 Variable Topics

We investigated the distribution of topics of the variables investigated in our literature sample, as represented by the variable categories defined in this work. Multiple categories can apply to a single variable, and this was the case for 138 out of 413 variables.

Of those categories referring to components of usability, e.g., related to either effectiveness, efficiency, or satisfaction [59], *Participant judgment* was the most used variable category (109 times). This may be the case since it applied to all variables representing some sort of participant judgment of an evaluated system or a task, e.g., preference [28], confidence in task correctness [2], or criticality of data [100]. This was followed by *security* (44), as a specific form of effectiveness, which is due to our sample focusing on DCUS. *Functionality* (24) as another form of effectiveness, *usability* overall (14) and *efficiency* (12) appeared less frequently.

Of the other variable categories, which were not related to usability, *participant characteristic* (93), *behavior* (69), and *experience* (47) were the most frequent. Examples for *participant characteristic* are type of participant, e.g., [65, 78], although there is a separate category specifically intended for this, demographics, like state of employment [76] or type of organization [74], and other characteristics relating to participants' opinions or attributes [16, 105]. *Experience*, e.g., with technology like programming languages [e.g., 28, 95], or specific tasks [e.g., 66, 75], often occurs together with *participant characteristic*, as experience can also be considered a defining characteristic of participants. In fact, the most frequent co-occurrence (30) between variable categories was between *experience* and *participant characteristic*. Variables measuring *behavior* included variables tracking participants' behavior during a study, e.g., the number of times they executed a program [54], number of visited websites [66], or lines of code submitted [114], and variables assigned to participants' outcomes retrospectively by researchers [e.g., 114]. Behavior could also be self-reported [e.g., 80, 105]. More specific categories, i.e., *participant type* (9) and *system type* (27) and *artifact related variables* (19) were not as frequent. There was a surprising amount of variables associated with meta-level aspects, such as *task related* (42) and *study related variables* (21), although some of these may be related to task success.

5.2 Use of Hypothesis Tests

We were especially interested in the frequency of hypothesis tests. In the studies in our sample, 7.30 hypothesis tests were conducted per study on average (SD=12.35), and given that a paper could encompass multiple studies, the number of hypothesis tests per paper ranged from 0 (for purely descriptive papers, like [37], or qualitative papers, like [53]) to 74 (M=8.6, SD=13.2). One of the studies, which the authors identified as qualitative, nevertheless contained 26 statistical hypothesis tests. In this comparison of end-users' and administrators' mental models of HTTPS, Fisher's exact tests were used to compare the appearance of various concepts in the mental models of the two participant types [65].

Since some studies included a large number of hypothesis tests, with outliers at 75, 51, and 30 hypothesis tests in a single study, we explored whether p-value correction methods were used in our sample. Figure 1 shows that the majority of studies (31/41) did not use corrections for any of the reported p-values. This likely includes some studies where these corrections were not necessary since the hypotheses tested were about different outcomes or were explicitly considered exploratory [101]. However, all three studies with the largest amount of tests did not include any corrections that we could identify for their hypothesis tests. This means that the results presented may be false positives.

The frequency of each of the different hypothesis tests within our sample is displayed in the left half of Figure 2. The most-used test (123 times) in our sample of papers was the non-parametric Wilcoxon rank-sum (a.k.a. Mann-Whitney U test), followed by two tests used for categorical data, the

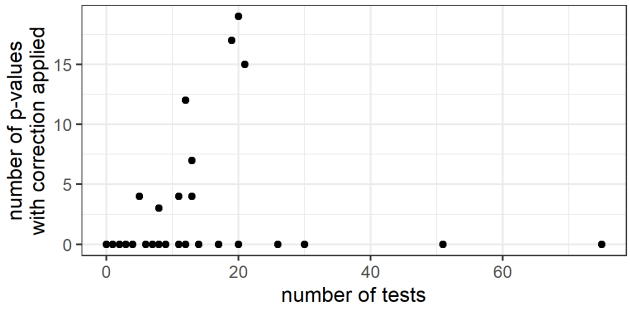


Figure 1: Scatterplot showing the frequency of corrected p-values in relation to non-corrected p-values per study

Fisher's exact and chi-squared Tests, which appeared 59 and 32 times respectively. We categorized the different hypothesis tests used in our sample according to the number and type of dependent and independent variables, the study design, and, in the case of categorical variables, the number of levels in the variable, see Table 1. All of the tests used only one dependent variable. In the remainder of our meta-analysis, we focus on those hypothesis tests with one categorical independent variable with two levels.

5.3 Completeness of Statistical Reporting

Ideally, each of these hypothesis tests would be reported in sufficient detail to be able to conduct power analysis using the data reported in the paper. However, this is not the case, as shown in Figure 3. Of those tests for which we can make this classification, the paired t-test was reported with sufficient information in all cases. However, it was reported only once. For the other tests, the completeness of reporting varied between 11.1% and 74.2% of sufficient reporting (mean=47.5%, sd=28.2%). The most frequently reported test, the Wilcoxon rank-sum test, was reported with sufficient information 38.2% of the time, or in 47 of 123 cases. Overall, this shows that reporting practices, even for these simple tests, are not sufficient to do power analysis using them as a basis about half of the time.

5.3.1 Power Meta-Analysis

To assess whether the field of DCUS suffers from underpowered studies similar to other fields as mentioned in section 2.2, we conducted a power meta-analysis based on Ellis [35, p.74]. This should not be confused with a post hoc power analysis since we did not use the effect sizes or p values reported in the papers. Instead, we only used the reported sample sizes and used G*Power to calculate the power to detect small, medium, and large effects (Cohen's d equivalent of 0.2, 0.5, and 0.8 [22]). This was possible for five of our seven types of hypothesis tests. We excluded Fisher's exact tests and McNemar's tests from this analysis since G*Power required input of effect size based on concrete data from the study, i.e., success proportions for the Fisher's exact test and the total proportion of discordant pairs for McNemar's test. Since post hoc power analysis using values directly from studies like this is not useful [35], we concentrated on the other five types of tests. We set $\alpha = 0.05$ for all analyses and assumed two-tailed tests. For the non-parametric tests, we used the minimal A.R.E. setting in G*Power to get a conservative estimate of the achieved power. Next, we calculated an average over all achieved power values at each level of effect size per included study, and then an overall average [35]. We considered a power of 0.8 to be the lower bound of what should commonly be aimed for [22][3].

Overall, our database contained 20 studies that reported enough statistical data to do this meta-analysis for at least one test. We found that nine of these had sufficient mean and median power to detect large effects, one had sufficient mean, and two had sufficient median power to detect medium effects and none of the studies had sufficient mean or median power to detect small effects. Conversely, eleven of the studies did not have 0.8 power to detect even large effects. However, over all the studies, the mean power to detect large effects was only slightly lower than the 0.8 we considered sufficient (mean=0.743, median=0.773). The mean power to detect medium effects was 0.455 (median=0.396), and for small effects, it was 0.132 (median=0.104). This again highlights the importance of a priori power analysis since developer studies are complex and resource intensive to run, and many do not have sufficient power under the common assumptions of $\alpha = 0.05$ and power of 0.8.

6 Systematic A Priori Power Analysis

Conducting an a priori power analysis requires researchers to know or guess the standardized effect sizes (e.g., Cohen's d) they expect or want to detect. Alternatively, they can also know or guess a non-standardized effect size (such as 1 point on a 7-point scale) and additional information, such as the standard deviation of the groups. In a mature field with many studies examining similar variables (e.g., blood pressure), expected effect sizes might be common knowledge. However, in the absence of closely related prior work, as is often the case in relatively new fields, such as DCUS, a more realistic approach is for the researcher to decide what the smallest practically relevant effect size is, that they want to be able to detect [35], also called the smallest effect size of interest. This can be determined by theoretical considerations but also by juxtaposing the benefit of the desired outcome with the costs to achieve this outcome or by considering practical limitations, such as the number of available participants [67]. In any case, researchers need experience and deep knowledge of their field

[3] Although deviations from this are perfectly fine when done consciously.

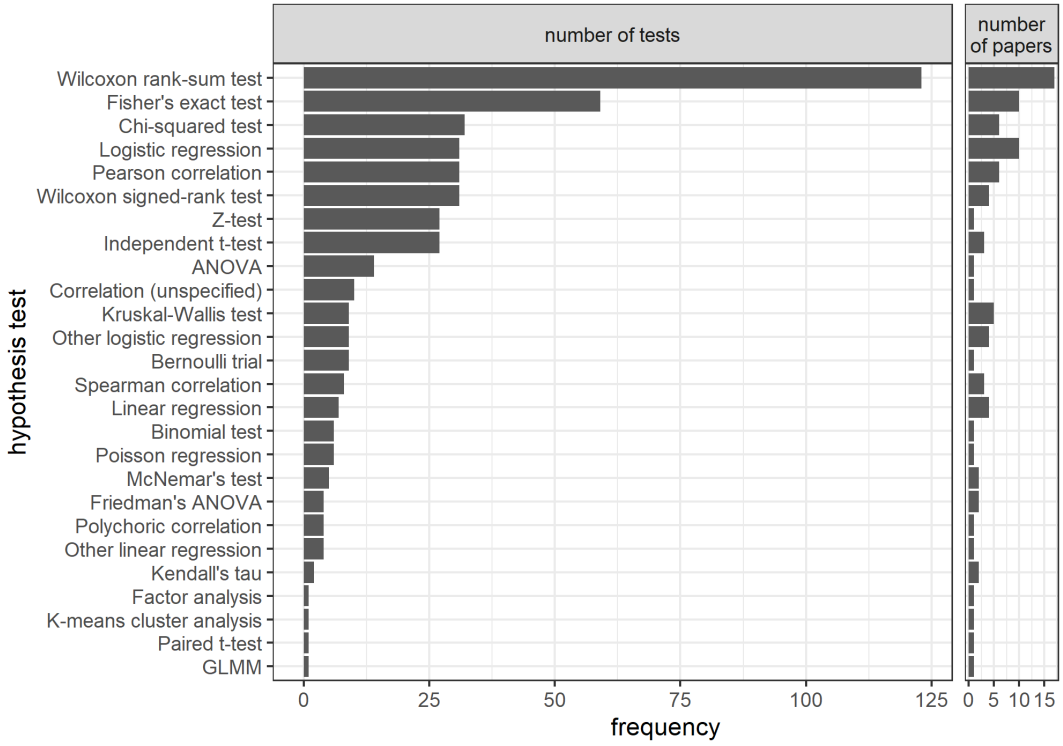


Figure 2: Left side: Frequency of type of hypothesis test in the sample, Right side: Number of papers using this hypothesis test in the sample

DVs in test	IVs in test	Participants	Test	ES	Descriptive stats	Paper
Attempted security (yes, no) security, behavior "participants who attempted to store user passwords securely, but struggled and then deleted their attempts from their solutions (this was coded as attempted but failed, or ABF). [...]"	Priming (priming, non-priming) study-related variable manipulated in experiment. "Priming - Participants were explicitly told to store the user passwords securely in the Introductory Text and in the Task Description."	computer science students (N=40)	FET	OR=19.02; d=1.62	Proportion p1=0.7 (priming); Proportion p2=0.1 (non-priming)	[76]
Secure (secure, insecure) security "In addition, we used a binary variable called secure which was given if participants used at least a hash function in their final solutions and thus did not store the passwords in plain text."	Warning displayed (yes, no) system type, study related variable whether a warning was displayed (could only happen in PyCrypto patch condition). "PyCrypto control condition, or the PyCrypto patch condition, where we tested our security warning"	Python developers (N=53)	FET	OR=56; d=2.22	Proportion p1=0.727 (yes); Proportion p2=0.0455 (no)	[47]

Table 2: Examples of information from the database which can be used for power analysis. The two examples are tests, where there was sufficient data to conduct power analysis. ES=effect size, FET=Fisher's exact test

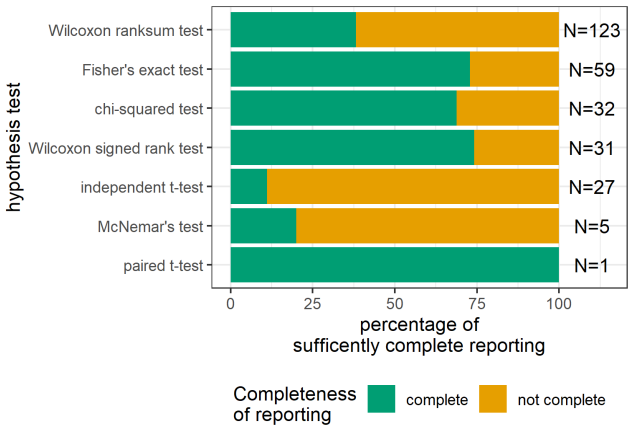


Figure 3: Stacked Barchart depicting the proportions sufficiently complete reporting for those tests at the focus of our analysis

to estimate effect sizes and this is one aspect that makes power analysis difficult [23]. In addition, used effect sizes [35] and power analysis procedures vary between different types of statistical tests [38].

Our systematization of knowledge aims to ease this process. Ideally, related studies have already been published, and standardized effect sizes are reported or can be calculated. In this case, the researcher needs to be able to find them. For this, they can query our database introduced in Section 4. Some types of information systematized in the database which are helpful for the search include variable categories, variables themselves and participant sample type. The participant sample type can help researchers identify prior work with a similar demographic to their planned study.

Table 2 shows an excerpt of the data which can be returned by the database. The first row contains all the data needed to perform a power analysis for a Fisher's exact test. A researcher could find it by any of the keywords or categories listed. The second row contains all the information to perform a power calculation for a different Fisher's exact test. The full data set is available on our companion website [4]. For each test, variables, effect sizes, relevant information to calculate them, the participant sample, and the source paper are listed and sorted by variable category.

Ideally, there will be several similar studies in the database, on which researchers can then base their effect size estimates on. However, this is currently unlikely since the field is still very young and diverse. But even if this is not the case, finding results for some of the variables of interest or a specific demographic can help refine effect size estimates.

[4] https://powerdb.info

6.1 Effectsize Meta-Analysis

As a final step in our systematization, we conducted a meta-analysis of the standardized effect sizes from tests where we had enough information for power analysis. With this, we aimed at providing a broad overview of effect sizes in DCUS which can be used to sanity check power analyses. While it is preferable to find exact or at least close matches, we also want to support researchers where this is not possible. Without related work, researchers basically have to guess standardized effect sizes or things like expected proportions or standard deviations. To at least give a frame of reference against which to judge these guesses, we examined the range of effect sizes present in the field of DCUS. We used the categories from Section 5.1 to aggregate the data for our guide.

To enable comparison and aggregation, we used the effect sizes directly reported in the paper where possible and converted them to Cohen's d, as this is one of the most widely used effect sizes in our sample. In other cases, we first used the provided data to calculate an effect size, which we then converted to Cohen's d. For converting d to odds ratio (OR), we used the formula from Haddock et al. [51], and a correction factor of 1.09, which is an average over the correction factors Poom and af Wåhlberg recommend for sample sizes between 20 and 100 [89], since developer studies mostly feature smaller sample sizes. To convert between d and ϕ, we used Rosenthal's formula [92] as described by Burns et al. [18]. Finally, as described above, sufficient data was not reported for many tests, and we exclude such tests from our analysis.

When judging the size of effects, effect sizes of Cohen's d=0.2 are generally regarded as small, d=0.5 as medium, and d=0.8 as large effect sizes [22]. When converted to other effect sizes, this yields OR=1.37, 2.21, 3.57 as small, medium, and large effect sizes displayed as odds ratios, ϕ=0.10, 0.24, 0.37 for the effect size ϕ used with χ^2-tests.

While we did not encounter the correlation coefficient r used as an effect size in this analysis, we nevertheless note for researchers encountering r, that effect sizes of r=0.1, 0.3, and 0.5 are considered small, medium and large effects respectively [22].

In our DCUS sample, the reported effect sizes ranged from equivalents of Cohen's d=0.004 to Cohen's d=2.22 (M=0.55, median=0.47, SD=0.42). We excluded two effect sizes from a Fisher's exact test in [76] and a McNemar's test in [106], where the reported success proportion in one of the groups was zero, and thus the effect size approaches infinity. Figure 4 shows the distributions of effect sizes separately for the variable categories of the dependent variables. So if researchers have to make educated guesses for their power analysis, they can compare their values with the violin plots to judge where in the spectrum they lie.

Three variable categories were not assigned to dependent variables with a present effect size: experience, system

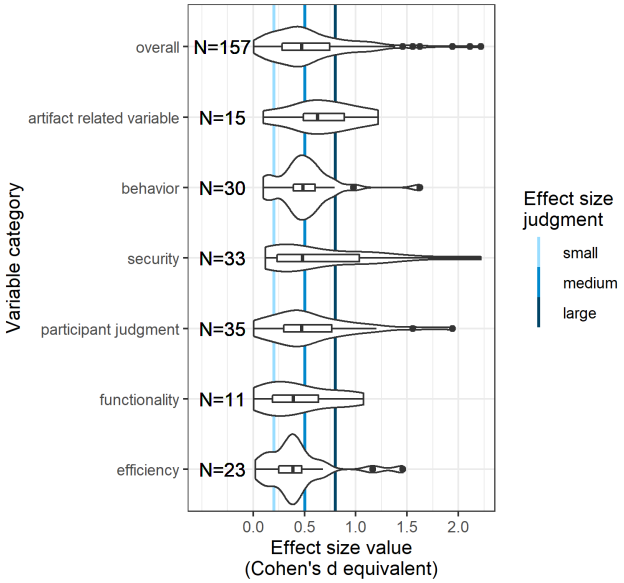


Figure 4: Violinplots for the distribution of effect sizes for tests, faceted by the variable categories of the dependent variable

type, and participant type. These categories were more frequently applied to independent variables. We did not plot data for variable categories where fewer than five effect sizes were reported. This is the case for four variable categories: usability (N=2), participant characteristics (N=1), task related (N=1), and study related variables (N=3). All of the five remaining variable categories exhibit a large variance of effect sizes, which range from negligible and small to large and very large effects. Median effects for tests with artifact-related variables as the dependent variable are in the medium range, and for all other categories and overall, the median effects are in the small range. However, in the cases of behavior, security, and participant judgment as well as overall, they border on medium according to Cohen [22]. We will update the online version when enough tests have been reported with the necessary statistical details.

6.2 Publication Bias Correction

There is one final important warning that needs to be highlighted. Irrespective of whether single entries from the database are used or our aggregation violin plots, researchers must be mindful of the effect of publication bias on reported effect sizes [58]. Since papers with statistically significant results are more likely to be published and studies in DCUS often have fairly poor power it should be expected that reported sample effect sizes are larger than true population effect sizes (see Ellis, p.79ff. [35]). Ideally, our database would also include work that is methodologically sound but did not get published due to statistically non-significant tests. However, since we could not think of any feasible way of including this at scale, we recommend taking this publication/sampling bias into account.

Consequently, when planning a study using effect sizes from related work or from our systematization or even a pre-study, it is recommendable to either correct the acquired effect size estimates [60, 111] or increase the desired sample size to be able to detect slightly smaller effects than ones reported in prior work. While this correction still requires some guesswork, it is a lot easier than having to guess blindly.

7 Power Analysis and Reporting Recommendations

While we hope that our systematization and database will already be a useful aid to researchers, it is still incomplete. A big hindrance in conducting our work was the fact that many tests were not reported with sufficient statistical information.

The American Psychological Association's publication manual contains a very comprehensive list of recommendations on how statistical tests should be reported [5]. Based on the APA and our findings in DCUS we want to highlight some recommendations for statistical reporting that we believe would be particularly helpful for future power analysis.

Report standardized effect sizes Wherever possible, report standardized effect sizes in addition to p values. Ellis provides an overview of effect sizes in Table 1.1 [35], and the guides in the appendix C have notes on effect sizes for seven commonly used hypothesis tests.

Report non-standardized effect sizes and descriptives Since standardized effect sizes can be unintuitive, also report descriptive statistics showing the effect size, e.g., there was a 1-point difference on a 7-point scale so that interested researchers can calculate effect sizes themselves. **Report frequencies** within all groups when comparing nominal variables, as effect size calculations commonly use these. **Report descriptive statistics for each group** of the independent variable(s) when comparing ordinal or interval variables. Report at least means, standard deviations, and group sizes. If this becomes too extensive, move this information to the appendix or supplemental material. **Make sure that the descriptives make sense**, e.g., a mean is not a good summary statistic for a bimodal distribution.

Make fully anonymized data sets available when needed If presenting all the descriptive statistics and frequencies is too extensive, e.g., for regression analyses with multiple independent variables, ideally, the anonymized data can be made available.

Hypothesis confirmation vs exploratory analysis State whether pre-defined hypotheses are being tested or

whether an exploratory analysis is being conducted [107].

8 Limitations

Our paper selection process focuses on the top-tier venues in which developer papers are published. Our sample does not include workshop papers or unpublished work, thus the publication and sampling bias needs to be taken into account as described above.

Additionally, even though descriptive data is useful when assessing effect sizes and conducting power analysis, we only included descriptive data in our database which was directly associated with a hypothesis test. The sheer amount of descriptive data reported in some papers, and the variety of ways it is reported, which included visualizations only partially enabling inference of exact numbers, tables, within the text or in the appendix, makes it hard to find a general structure for storing this type of data. We defer this to future work.

Our work focuses only on a priori power analysis as described in Section 2.2. Power, and as such, power analysis is situated within the NHST framework, and the analyses of the field and recommendations in this work apply within NHST. NHST results in point estimates about coefficients or effect sizes and a priori power analysis serves to determine the number of participants needed to detect such an effect at a specified significance level. The practice of focusing on the value of point estimates, rather than the precision of the estimates has also been criticized [69]. Even when planning a study with adequate power to detect an effect, the confidence intervals around it can be wide, leading to little precision of the effect size. Different analysis methods also exist, e.g., Bayesian analysis [63, 64]. Additionally, some statistical analyses, like regression, are not used merely in NHST to falsify hypotheses, but also to make predictions, and different judgment criteria would apply in these cases [19]. In predictive analysis, i.e., what is often known as machine learning, the influence and explanatory power of individual variables included in the model is not as important as the accuracy of the prediction [17, 19]. Other criteria for what constitutes good reporting may apply in these cases than what we have covered in this work. However, NHST is commonly used in DCUS specifically and HCI in general. We did not in fact encounter any Bayesian analyses in our sample of analyzed DCUS papers and given that recruiting software developers is hard [4], having sufficient data for large-scale predictive analyses is likely rare in this field. In conclusion, we believe that our contributions align with common methods used in DCUS at the time.

Finally, we only conducted our meta-analysis on tests with a single categorical IV and a single DV. Thus, our work is limited to assisting in the power analysis of the following tests: Fisher's exact test, chi-squared test, McNemar's test, Wilcoxon rank-sum tests, Wilcoxon signed-rank tests, and paired and independent t-tests. We will extend this in future work. However, in combination with simplifications even the current data might offer benefits for the assessment of more complex tests[68, 87].

9 Conclusion

In this work, we systematized 467 tests and 413 variables from a data set of 54 DCUS papers published in top-tier venues. We examined their methodology and reporting of statistical results, as well as their power to detect effects of different sizes, which was not sufficient for small effects, and only sufficient for large effects in about half of the papers, where enough information was reported to analyze the power. We provide domain-specific effect size ranges for different categories of variables, which can serve as a fall-back for effect size estimation in a priori power analysis, with effect sizes in DCUS averaging at Cohen's d=0.55, when considering all variable categories. The raw data on which these ranges are based is included in a searchable database of extracted information from these papers, which other researchers can use to reproduce our analyses and facilitate their own power analyses. The brief guides in the appendix can supply further assistance in conducting power analysis for some simple but often-used hypothesis tests. When reporting statistical results, authors should include effect sizes, both standardized and non-standardized, as well as descriptive statistics for each condition as a way to foster the use of power analysis in sample size planning and to enable the re-analysis of results through meta-analysis.

10 Future Work

As stated in the limitations section, we currently only cover a subset of all tests. In future work, we plan to extend this list to include more tests. The current implementation of our database does not have a custom user interface and is operated using SQL or other query tools to access and understand the data within. Ongoing work aims to improve the usability of querying the database. We also plan to extend our approach to the whole field of usable security and privacy. While DCUS faces particular challenges when recruiting participants, we believe end-user studies would also benefit from a priori power analysis. To aid in this extension and general upkeep we plan on developing a public-facing interface to the database so researchers can add their own papers. In the future, the database could also be used to explore other aspects of methodology use, such as the number of coders and use of inter-rater reliability in qualitative work, whether the behavior is measured objectively or subjectively using the categorization of variables, or to conduct sensitivity analysis, i.e., to analyze the power to detect effects.

Acknowledgments

We would like to thank all the students working on paper categorization or the database application in various courses between 2021 and 2023: Ahmad Alaya, Somar Aljabr, Ahmad Assaf, Marwan Jaradat, Tansen Khan, Heng-yi Lin, Florin Martius and Atacan Süder, our colleagues Lisa Geierhaas, Eva Gerlitz, Maximilian Häring, Mischa Meier, Stephan Plöger, and Klaus Tulbure, who aided with paper collection and selection, as well as Simon Bong and Theo Raimbault, who transferred information from collected papers into the database.

References

[1] Yasemin Acar, Michael Backes, Sascha Fahl, Simson Garfinkel, Doowon Kim, Michelle L. Mazurek, and Christian Stransky. Comparing the Usability of Cryptographic APIs. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 154–171, San Jose, CA, USA, May 2017. IEEE.

[2] Yasemin Acar, Michael Backes, Sascha Fahl, Doowon Kim, Michelle L. Mazurek, and Christian Stransky. You Get Where You're Looking for: The Impact of Information Sources on Code Security. In *2016 IEEE Symposium on Security and Privacy (SP)*, pages 289–305, San Jose, CA, May 2016. IEEE.

[3] Yasemin Acar, Sascha Fahl, and Michelle L. Mazurek. You are Not Your Developer, Either: A Research Agenda for Usable Security and Privacy Research Beyond End Users. In *2016 IEEE Cybersecurity Development (SecDev)*, pages 3–8, Boston, MA, USA, November 2016. IEEE.

[4] Yasemin Acar, Christian Stransky, Dominik Wermke, Michelle L. Mazurek, and Sascha Fahl. Security Developer Studies with {GitHub} Users: Exploring a Convenience Sample. In *Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017)*, pages 81–95, 2017.

[5] American Psychological Association, editor. *Publication Manual of the American Psychological Association*. American Psychological Association, Washington, DC, seventh edition edition, 2020.

[6] Hala Assal and Sonia Chiasson. Security in the Software Development Lifecycle. In *Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018)*, SOUPS '18, pages 281–296, Baltimore, MD, August 2018. USENIX Association.

[7] Donald R. Atkinson, Michael J. Furlong, and Bruce E. Wampold. Statistical significance, reviewer evaluations, and the scientific process: Is there a (statistically) significant relationship? *Journal of Counseling Psychology*, 29(2):189–194, 1982.

[8] Peter C. Austin and Ewout W. Steyerberg. The number of subjects per variable required in linear regression analyses. *Journal of Clinical Epidemiology*, 68(6):627–636, June 2015.

[9] Nathaniel Ayewah and William Pugh. A report on a survey and study of static analysis users. In *Proceedings of the 2008 Workshop on Defects in Large Software Systems*, DEFECTS '08, pages 1–5, New York, NY, USA, July 2008. Association for Computing Machinery.

[10] Thom Baguley. Standardized or simple effect size: What should be reported? *British Journal of Psychology*, 100(3):603–617, 2009.

[11] Khadija Baig, Elisa Kazan, Kalpana Hundlani, Sana Maqsood, and Sonia Chiasson. Replication: Effects of Media on the Mental Models of Technical Users. In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*, pages 119–138, 2021.

[12] Jack J. Baroudi and Wanda J. Orlikowski. The problem of statistical power in MIS research. *MIS Quarterly*, 13(1):87–106, 1989.

[13] Dirk Bassler, Matthias Briel, Victor M. Montori, Melanie Lane, Paul Glasziou, Qi Zhou, Diane Heels-Ansdell, Stephen D. Walter, Gordon H. Guyatt, and and the STOPIT-2 Study Group. Stopping Randomized Trials Early for Benefit and Estimation of Treatment Effects: Systematic Review and Meta-regression Analysis. *JAMA*, 303(12):1180–1187, March 2010.

[14] Scott Bezeau and Roger Graves. Statistical Power and Effect Sizes of Clinical Neuropsychology Research. *Journal of Clinical and Experimental Neuropsychology*, 23(3):399–406, June 2001.

[15] Robert Biddle, Sonia Chiasson, and P.C. Van Oorschot. Graphical passwords: Learning from the first twelve years. *ACM Computing Surveys*, 44(4):19:1–19:41, September 2012.

[16] Larissa Braz, Enrico Fregnan, Gül Çalikli, and Alberto Bacchelli. Why Don't Developers Detect Improper Input Validation? '; DROP TABLE Papers; –. In *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)*, pages 499–511, May 2021.

[17] Leo Breiman. Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3):199–231, August 2001.

[18] Matthew K. Burns, Anne F. Zaslofsky, Rebecca Kanive, and David C. Parker. Meta-Analysis of Incremental Rehearsal Using Phi Coefficients to Compare Single-Case and Group Designs. *Journal of Behavioral Education*, 21(3):185–202, September 2012.

[19] Danilo Bzdok, Denis Engemann, and Bertrand Thirion. Inference and Prediction Diverge in Biomedicine. *Patterns*, 1(8):100119, November 2020.

[20] Kelly Caine. Local standards for sample size at chi. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, page 981–992, New York, NY, USA, 2016. Association for Computing Machinery.

[21] Andy Cockburn, Pierre Dragicevic, Lonni Besançon, and Carl Gutwin. Threats of a Replication Crisis in Empirical Computer Science. *Communications of the ACM*, 63(8):70–79, 2020.

[22] Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. L. Erlbaum Associates, Hillsdale, N.J, 2nd ed edition, 1988.

[23] Jacob Cohen. A power primer. In A. E. Kazdin, editor, *Methodological Issues and Strategies in Clinical Research*, pages 279–284. American Psychological Association, Washington, DC, US, 2016.

[24] Shaanan Cohney, Ross Teixeira, Anne Kohlbrenner, Arvind Narayanan, Mihir Kshirsagar, Yan Shvartzshnaider, and Madelyn Sanfilippo. Virtual Classrooms and Real Harms: Remote Learning at {U.S}. Universities. In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*, pages 653–674, 2021.

[25] David A. Cook and Rose Hatala. Got power? A systematic review of sample size adequacy in health professions education research. *Advances in Health Sciences Education*, 20(1):73–83, March 2015.

[26] Anastasia Danilova, Alena Naiakshina, Johanna Deuter, and Matthew Smith. Replication: On the Ecological Validity of Online Security Developer Studies: Exploring Deception in a Password-Storage Study with Freelancers. In *Sixteenth Symposium on Usable Privacy and Security (SOUPS 2020)*, pages 165–183. USENIX Association, August 2020.

[27] Anastasia Danilova, Alena Naiakshina, Anna Rasgauski, and Matthew Smith. Code Reviewing as Methodology for Online Security Studies with Developers - A Case Study with Freelancers on Password Storage. In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*, pages 397–416, 2021.

[28] Anastasia Danilova, Alena Naiakshina, and Matthew Smith. One Size Does Not Fit All: A Grounded Theory and Online Survey Study of Developer Preferences for Security Warning Types. In *2020 IEEE/ACM 42nd International Conference on Software Engineering (ICSE)*, pages 136–148, October 2020.

[29] Erik Derr, Sven Bugiel, Sascha Fahl, Yasemin Acar, and Michael Backes. Keep me Updated: An Empirical Study of Third-Party Library Updatability on Android. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, CCS '17, pages 2187–2200, Dallas Texas USA, October 2017. ACM.

[30] Julian di Stephano. How much power is enough? Against the development of an arbitrary convention for statistical power calculations. *Functional Ecology*, 17(5):707–709, 2003.

[31] Constanze Dietrich, Katharina Krombholz, Kevin Borgolte, and Tobias Fiebig. Investigating System Operators' Perspective on Security Misconfigurations. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, CCS '18, pages 1272–1289, New York, NY, USA, October 2018. Association for Computing Machinery.

[32] Verena Distler, Matthias Fassl, Hana Habib, Katharina Krombholz, Gabriele Lenzini, Carine Lallemand, Lorrie Faith Cranor, and Vincent Koenig. A Systematic Literature Review of Empirical Methods and Risk Representation in Usable Privacy and Security Research. *ACM Transactions on Computer-Human Interaction*, 28(6):43:1–43:50, December 2021.

[33] Alexander Eiselmayer. Supporting the Design and Analysis of HCI Experiments. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–8, Honolulu HI USA, April 2020. ACM.

[34] Alexander Eiselmayer, Chat Wacharamanotham, Michel Beaudouin-Lafon, and Wendy E. Mackay. Touchstone2: An interactive environment for exploring trade-offs in HCI experiment design. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 1–11, New York, NY, USA, 2019. Association for Computing Machinery.

[35] Paul D. Ellis. *The Essential Guide to Effect Sizes. Statistical Power, Meta-Analysis, and the Interpretation of Research Results*. Cambridge University Press, 2010.

[36] Connor Esterwood, Kyle Essenmacher, Han Yang, Fanpan Zeng, and Lionel Peter Robert. A Meta-Analysis of Human Personality and Robot Acceptance in Human-Robot Interaction. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, pages 1–18, New York, NY, USA, May 2021. Association for Computing Machinery.

[37] Sascha Fahl, Marian Harbach, Henning Perl, Markus Koetter, and Matthew Smith. Rethinking SSL development in an appified world. In *Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security - CCS '13*, CCS '13, pages 49–60, Berlin, Germany, 2013. ACM Press.

[38] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39(2):175–191, May 2007.

[39] Andy Field, Jeremy Miles, and Zoe Field. *Discovering Statistics Using R*. SAGE Publications Ltd, London ; Thousand Oaks, Calif, 1. edition edition, April 2012.

[40] Felix Fischer, Yannick Stachelscheid, and Jens Grossklags. The Effect of Google Search on Software Security: Unobtrusive Security Interventions via Content Re-ranking. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, CCS '21, pages 3070–3084, New York, NY, USA, November 2021. Association for Computing Machinery.

[41] Felix Fischer, Huang Xiao, Ching-Yu Kao, Yannick Stachelscheid, Benjamin Johnson, Danial Razar, Paul Fawkesley, Nat Buckley, Konstantin Bottinger, Paul Muntean, and Jens Grossklags. Stack Overflow Considered Helpful! Deep Learning Security Nudges Towards Stronger Cryptography. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 339–356, Santa Clara, CA, USA, August 2019. USENIX Association.

[42] Boris Freidlin and Edward L Korn. Stopping clinical trials early for benefit: Impact on estimation. *Clinical Trials*, 6(2):119–125, April 2009.

[43] Kelsey R Fulton, Anna Chan, Daniel Votipka, Michael Hicks, and Michelle L Mazurek. Benefits and Drawbacks of Adopting a Secure Programming Language: Rust as a Case Study. In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*, SOUPS '21, page 20. USENIX Association, August 2021.

[44] Eva Gerlitz, Maximilian Häring, and Matthew Smith. Please do not use !?_ or your License Plate Number: Analyzing Password Policies in German Companies. In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*, pages 17–36, 2021.

[45] Steven N. Goodman and Jesse A. Berlin. The Use of Predicted Confidence Intervals When Planning Experiments and the Misuse of Power When Interpreting Results. *Annals of Internal Medicine*, 121(3):200–206, 1994.

[46] Peter Leo Gorski, Yasemin Acar, Luigi Lo Iacono, and Sascha Fahl. Listen to Developers! A Participatory Design Study on Security Warnings for Cryptographic APIs. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, Honolulu HI USA, April 2020. ACM.

[47] Peter Leo Gorski, Luigi Lo Iacono, Dominik Wermke, Christian Stransky, Sebastian Moeller, Yasemin Acar, and Sascha Fahl. Developers Deserve Security Warnings, Too. In *Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018)*, SOUPS '18, pages 265–281, Baltimore, MD, August 2018. USENIX Association.

[48] Matthew Green and Matthew Smith. Developers are Not the Enemy!: The Need for Usable Security APIs. *IEEE Security & Privacy*, 14(5):40–46, September 2016.

[49] Samuel B. Green. How Many Subjects Does It Take To Do A Regression Analysis. *Multivariate Behavioral Research*, 26(3):499–510, July 1991.

[50] Thomas Groß. Fidelity of Statistical Reporting in 10 Years of Cyber Security User Studies. In Thomas Groß and Theo Tryfonas, editors, *Socio-Technical Aspects in Security and Trust*, Lecture Notes in Computer Science, pages 3–26, Cham, 2021. Springer International Publishing.

[51] C. Keith Haddock, David Rindskopf, and William R. Shadish. Using odds ratios as effect sizes for meta-analysis of dichotomous data: A primer on methods and issues. *Psychological Methods*, 3(3):339–353, 1998.

[52] Joseph Hallett, Nikhil Patnaik, Benjamin Shreeve, and Awais Rashid. “Do this! Do that!, and Nothing will Happen” Do Specifications Lead to Securely Stored Passwords? In *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)*, pages 486–498, May 2021.

[53] Julie M Haney, Mary F Theofanos, Yasemin Acar, and Sandra Spickard Prettyman. "We make it a big deal in the company": Security Mindsets in Organizations that Develop Cryptographic Products. In *Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018)*, SOUPS '18, pages 357–373, Baltimore, MD, August 2018. USENIX Association.

[54] Norman Hänsch, Andrea Schankin, and Mykolai Protsenko. Programming Experience Might Not Help in Comprehending Obfuscated Source Code Efficiently. In *Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018)*, pages 341–356, Baltimore, MD, August 2018. USENIX Association.

[55] Moonseong Heo, Singh R. Nair, Judith Wylie-Rosett, Myles S. Faith, Angelo Pietrobelli, Nancy R. Glassman, Sarah N. Martin, Stephanie Dickinson, and David B. Allison. Trial Characteristics and Appropriateness of Statistical Methods Applied for Design and Analysis of Randomized School-Based Studies Addressing Weight-Related Issues: A Literature Review. *Journal of Obesity*, 2018:e8767315, June 2018.

[56] John M Hoenig and Dennis M Heisey. The Abuse of Power. *The American Statistician*, 55(1):19–24, February 2001.

[57] Kasper Hornbæk and Effie Lai-Chong Law. Meta-analysis of correlations among usability measures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 617–626, San Jose California USA, April 2007. ACM.

[58] John P. A. Ioannidis. Why Most Discovered True Associations Are Inflated. *Epidemiology*, 19(5):640–648, 2008.

[59] ISO Central Secretary. Ergonomics of human-system interaction - Part 11: Usability: Definitions and concepts. Standard ISO 9241-11:2018, International Organization for Standardization, Geneva, CH, 2018.

[60] Andreas Ivarsson, Mark B. Andersen, Urban Johnson, and Magnus Lindwall. To adjust or not adjust: Nonparametric effect sizes, confidence intervals, and real-world meaning. *Psychology of Sport and Exercise*, 14(1):97–102, January 2013.

[61] Brittany Johnson, Yoonki Song, Emerson Murphy-Hill, and Robert Bowdidge. Why don't software developers use static analysis tools to find bugs? In *2013 35th International Conference on Software Engineering (ICSE)*, pages 672–681, May 2013.

[62] Vigdis By Kampenes, Tore Dybå, Jo E. Hannay, and Dag I.K. Sjøberg. A systematic review of effect size in software engineering experiments. *Information and Software Technology*, 49(11):1073–1086, 2007.

[63] Maurits Kaptein and Judy Robertson. Rethinking statistical analysis methods for CHI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 1105–1114, New York, NY, USA, May 2012. Association for Computing Machinery.

[64] Matthew Kay, Gregory L. Nelson, and Eric B. Hekler. Researcher-Centered Design of Statistics: Why Bayesian Statistics Better Fit the Culture and Incentives of HCI. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 4521–4532, San Jose California USA, May 2016. ACM.

[65] Katharina Krombholz, Karoline Busse, Katharina Pfeffer, Matthew Smith, and Emanuel von Zezschwitz. "If HTTPS Were Secure, I Wouldn't Need 2FA" - End User and Administrator Mental Models of HTTPS. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 246–263, San Francisco, CA, USA, May 2019. IEEE.

[66] Katharina Krombholz, Wilfried Mayer, Martin Schmiedecker, and Edgar Weippl. “I Have No Idea What I'm Doing” – On the Usability of Deploying HTTPS. In *26th USENIX Security Symposium (USENIX Security 17)*, pages 1339–1356, Vancouver, BC, Canada, 2017. USENIX Association.

[67] Daniël Lakens. Performing high-powered studies efficiently with sequential analyses. *European Journal of Social Psychology*, 44(7):701–710, 2014.

[68] Sean P. Lane and Erin P. Hennes. Power struggles: Estimating sample size for multilevel relationships research. *Journal of Social and Personal Relationships*, 35(1):7–31, January 2018.

[69] Scott E. Maxwell, Ken Kelley, and Joseph R. Rausch. Sample Size Planning for Statistical Power and Accuracy in Parameter Estimation. *Annual Review of Psychology*, 59(1):537–563, January 2008.

[70] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. Reliability and Inter-rater Reliability in Qualitative Research: Norms and Guidelines for CSCW and HCI Practice. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–23, November 2019.

[71] Abraham H Mhaidli, Yixin Zou, and Florian Schaub. "We Can't Live Without Them!" App Developers' Adoption of Ad Networks and Their Considerations of Consumer Risks. In *Fifteenth Symposium on Usable Privacy and Security (SOUPS 2019)*, pages 225–244, Santa Clara, CA, August 2019. USENIX Association.

[72] Victor M. Montori, P. J. Devereaux, Neill K. J. Adhikari, Karen E. A. Burns, Christoph H. Eggert, Matthias Briel, Christina Lacchetti, Teresa W. Leung, Elizabeth Darling, Dianne M. Bryant, Heiner C. Bucher, Holger J. Schünemann, Maureen O. Meade, Deborah J. Cook, Patricia J. Erwin, Amit Sood, Richa Sood, Benjamin Lo, Carly A. Thompson, Qi Zhou, Edward Mills, and Gordon H. Guyatt. Randomized Trials Stopped Early for BenefitA Systematic Review. *JAMA*, 294(17):2203–2209, November 2005.

[73] Sarah Nadi, Stefan Krüger, Mira Mezini, and Eric Bodden. Jumping through hoops: Why do Java developers struggle with cryptography APIs? In *Proceedings of the 38th International Conference on Software Engineering*, pages 935–946, Austin Texas, May 2016. ACM.

[74] Alena Naiakshina, Anastasia Danilova, Eva Gerlitz, and Matthew Smith. On Conducting Security Developer Studies with CS Students: Examining a Password-Storage Study with CS Students, Freelancers, and Company Developers. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, Honolulu HI USA, April 2020. ACM.

[75] Alena Naiakshina, Anastasia Danilova, Eva Gerlitz, Emanuel von Zezschwitz, and Matthew Smith. "If you want, I can store the encrypted password": A Password-Storage Field Study with Freelance Developers. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 1–12, Glasgow, Scotland, Uk, May 2019. ACM.

[76] Alena Naiakshina, Anastasia Danilova, and Christian Tiefenau. Deception Task Design in Developer Password Studies: Exploring a Student Sample. In *Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018)*, SOUPS '18, pages 297–313, Baltimore, MD, August 2018. USENIX Association.

[77] Alena Naiakshina, Anastasia Danilova, Christian Tiefenau, Marco Herzog, Sergej Dechand, and Matthew Smith. Why Do Developers Get Password Storage Wrong?: A Qualitative Usability Study. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, CCS '17, pages 311–328, Dallas Texas USA, October 2017. ACM.

[78] Duc Cuong Nguyen, Dominik Wermke, Yasemin Acar, Michael Backes, Charles Weir, and Sascha Fahl. A Stitch in Time: Supporting Android Developers in Writing Secure Code. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, CCS '17, pages 1065–1077, Dallas Texas USA, October 2017. ACM.

[79] Jakob Nielsen and Jonathan Levy. Measuring usability: Preference vs. performance. *Communications of the ACM*, 37(4):66–75, April 1994.

[80] Tim Nosco, Jared Ziegler, and Zechariah Clark. The Industrial Age of Hacking. In *29th USENIX Security Symposium (USENIX Security 20)*, USENIX Security '20, pages 1129–1146. USENIX Association, August 2020.

[81] Natalia Obukhova. A Meta-Analysis of Effect Sizes of CHI Typing Experiments. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI EA '21, pages 1–7, New York, NY, USA, May 2021. Association for Computing Machinery.

[82] Daniela Seabra Oliveira, Tian Lin, Muhammad Sajidur Rahman, Rad Akefirad, Donovan Ellis, Eliany Perez, and Rahul Bobhate. API Blindspots: Why Experienced Developers Write Vulnerable Code. In *Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018)*, SOUPS '18, pages 315–328, Baltimore, MD, August 2018. USENIX Association.

[83] Marten Oltrogge, Yasemin Acar, Sergej Dechand, Matthew Smith, and Sascha Fahl. To Pin or Not to Pin Helping App Developers Bullet Proof Their TLS Connections. In *24th 24th USENIX Security Symposium (USENIX Security 15*, pages 239–254, Washington, D.C., USA, August 2015. USENIX Association.

[84] Anna-Marie Ortloff, Matthias Fassl, Alexander Ponticello, Florin Martius, Anne Mertens, Katharina Krombholz, and Matthew Smith. Different researchers, different results? analyzing the influence of researcher experience and data type during qualitative analysis of an interview and survey study on security advice. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery.

[85] Hernan Palombo, Armin Ziaie Tabari, Daniel Lende, Jay Ligatti, and Xinming Ou. An Ethnographic Understanding of Software (In)Security and a Co-Creation Model to Improve Secure Software Development. In *Sixteenth Symposium on Usable Privacy and Security (SOUPS 2020)*, SOUPS '20, page 17. USENIX Association, August 2020.

[86] Ivan Pashchenko, Duc-Ly Vu, and Fabio Massacci. A Qualitative Study of Dependency Management and Its Security Implications. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, pages 1513–1531, Virtual Event USA, October 2020. ACM.

[87] Marco Perugini, Marcello Gallucci, and Giulio Costantini. A Practical Primer To Power Analysis for Simple Experimental Designs. 31(1):20, July 2018.

[88] Stephan Plöger, Mischa Meier, and Matthew Smith. A Qualitative Usability Evaluation of the Clang Static Analyzer and {libFuzzer} with {CS} Students and {CTF} Players. In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*, pages 553–572, 2021.

[89] Leo Poom and Anders af Wåhlberg. Accuracy of conversion formula for effect sizes: A Monte Carlo simulation. *Research Synthesis Methods*, 13(4):508–519, 2022.

[90] Lutz Prechelt. Plat_Forms: A Web Development Platform Comparison by an Exploratory Experiment Searching for Emergent Platform Properties. *IEEE Transactions on Software Engineering*, 37(1):95–108, January 2011.

[91] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022.

[92] Robert Rosenthal. Parametric Measures of Effect Size. In Harris Cooper and Larry Hedges, editors, *The Handbook of Research Synthesis*, pages 231–244. Russel Sage Foundation, New York, 1994.

[93] Joseph Rossi. Statistical Power of Psychological Research: What Have We Gained in 20 Years? *Journal of Consulting and Clinical Psychology*, 58(5):646–656, 1990.

[94] Sebastian Roth, Lea Gröber, Michael Backes, Katharina Krombholz, and Ben Stock. 12 Angry Developers - A Qualitative Study on Developers' Struggles with CSP. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, CCS '21, pages 3085–3103, New York, NY, USA, November 2021. Association for Computing Machinery.

[95] Andrew Ruef, Michael Hicks, James Parker, Dave Levin, Michelle L. Mazurek, and Piotr Mardziel. Build It, Break It, Fix It: Contesting Secure Development. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 690–703, Vienna Austria, October 2016. ACM.

[96] Roberta W Scherer, Joerg J Meerpohl, Nadine Pfeifer, Christine Schmucker, Guido Schwarzer, and Erik von Elm. Full publication of results initially presented in abstracts. *Cochrane Database of Systematic Reviews*, 2018(11), November 2018.

[97] I. Manjula Schou and Ian C. Marschner. Meta-analysis of clinical trials with early stopping: An investigation of potential bias. *Statistics in Medicine*, 32(28):4859–4874, 2013.

[98] Skipper Seabold and Josef Perktold. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*, 2010.

[99] Peter Sedlmeier and Gerd Gigerenzer. Do studies of statistical power have an effect on the power of studies? In *Methodological Issues & Strategies in Clinical Research*, pages 389–406. American Psychological Association, Washington, DC, US, 1992.

[100] Swapneel Sheth, Gail Kaiser, and Walid Maalej. Us and them: A study of privacy requirements across north america, asia, and europe. In *Proceedings of the 36th International Conference on Software Engineering*, ICSE 2014, pages 859–870, New York, NY, USA, May 2014. Association for Computing Machinery.

[101] David L. Streiner and Geoffrey R. Norman. Correction for Multiple Testing: Is There a Resolution? *CHEST*, 140(1):16–18, July 2011.

[102] Mohammad Tahaei, Alisa Frik, and Kami Vaniea. Deciding on Personalized Ads: Nudging Developers About User Privacy. In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*, pages 573–596. USENIX Association, 2021.

[103] Mohammad Tahaei and Kami Vaniea. A Survey on Developer-Centred Security. In *2019 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, pages 129–138, June 2019.

[104] Meinald T. Thielsch, Jana Scharfen, Ehsan Masoudi, and Meike Reuter. Visual Aesthetics and Performance: A First Meta-Analysis. In *Proceedings of Mensch Und Computer 2019*, pages 199–210, Hamburg Germany, September 2019. ACM.

[105] Christian Tiefenau, Maximilian Häring, and Katharina Krombholz. Security, Availability, and Multiple Information Sources: Exploring Update Behavior of System Administrators. In *Sixteenth Symposium on Usable Privacy and Security (SOUPS 2020)*, SOUPS '20, pages 239–258. USENIX Association, August 2020.

[106] Christian Tiefenau, Emanuel von Zezschwitz, Maximilian Häring, Katharina Krombholz, and Matthew Smith. A Usability Evaluation of Let's Encrypt and Certbot: Usable Security Done Right. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, CCS '19, pages 1971–1988, London United Kingdom, November 2019. ACM.

[107] Andrew T. Tredennick, Giles Hooker, Stephen P. Ellner, and Peter B. Adler. A practical guide to selecting models for exploration, inference, and prediction in ecology. *Ecology*, 102(6), June 2021.

[108] Patrizio E. Tressoldi and David Giofré. The pervasive avoidance of prospective statistical power: Major consequences and practical solutions. *Frontiers in Psychology*, 6, 2015.

[109] Patrizio E. Tressoldi, David Giofré, Francesco Sella, and Geoff Cumming. High Impact = High Statistical Standards? Not Necessarily So. *PLOS ONE*, 8(2):e56180, February 2013.

[110] Anwesh Tuladhar, Daniel Lende, Jay Ligatti, and Xinming Ou. An Analysis of the Role of Situated Learning in Starting a Security Culture in a Software Company. In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*, pages 617–632. USENIX Association, 2021.

[111] Tammi Vacha-Haase and Bruce Thompson. How to Estimate and Interpret Various Effect Sizes. *Journal of Counseling Psychology*, 51(4):473–481, 2004.

[112] Dirk van der Linden, Pauline Anthonysamy, Bashar Nuseibeh, Thein Than Tun, Marian Petre, Mark Levine, John Towse, and Awais Rashid. Schrödinger's Security: Opening the Box on App Developers' Security Rationale. In *2020 IEEE/ACM 42nd International Conference on Software Engineering (ICSE)*, pages 149–160, October 2020.

[113] Ivan Vankov, Jeffrey Bowers, and Marcus Munafò. On the persistence of low power in psychological science. *Q J Exp Psychol (Hove)*, 67(6):1037–1040, 2014.

[114] Daniel Votipka, Kelsey R Fulton, James Parker, Matthew Hou, Michelle L Mazurek, and Michael Hicks. Understanding security mistakes developers make: Qualitative analysis from Build It, Break It, Fix It. In *29th USENIX Security Symposium (USENIX Security 20)*, USENIX Security '20, pages 109–126. USENIX Association, August 2020.

[115] Daniel Votipka, Seth Rabin, Kristopher Micinski, Jeffrey S Foster, and Michelle L Mazurek. An Observational Investigation of Reverse Engineers' Processes. In *29th USENIX Security Symposium (USENIX Security 20)*, pages 1875–1892. USENIX Association, August 2020.

[116] Daniel Votipka, Rock Stevens, Elissa Redmiles, Jeremy Hu, and Michelle Mazurek. Hackers vs. Testers: A Comparison of Software Vulnerability Discovery Processes. In *2018 IEEE Symposium on Security and Privacy (SP)*, pages 374–391, San Francisco, CA, May 2018. IEEE.

[117] Charles Weir, Ben Hermann, and Sascha Fahl. From Needs to Actions to Secure Apps? The Effect of Requirements and Developer Practices on App Security. In *29th USENIX Security Symposium (USENIX Security 20)*, pages 289–305. USENIX Association, August 2020.

[118] Charles Weir, Awais Rashid, and James Noble. How to improve the security skills of mobile app developers? Comparing and contrasting expert views. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*, SOUPS'16, Denver Colorado USA, June 2016. USENIX Association.

[119] Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686, 2019.

[120] Khaled Yakdan, Sergej Dechand, Elmar Gerhards-Padilla, and Matthew Smith. Helping Johnny to Analyze Malware: A Usability-Optimized Decompiler and Malware Analysis User Study. In *2016 IEEE Symposium on Security and Privacy (SP)*, pages 158–177, San Jose, CA, USA, May 2016. IEEE.

[121] Miuyin Yong Wong, Matthew Landen, Manos Antonakakis, Douglas M. Blough, Elissa M. Redmiles, and Mustaque Ahamad. An Inside Look into the Practice of Malware Analysis. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, CCS '21, pages 3053–3069, New York, NY, USA, November 2021. Association for Computing Machinery.

[122] Koen Yskout, Riccardo Scandariato, and Wouter Joosen. Does organizing security patterns focus architectural choices? In *2012 34th International Conference on Software Engineering (ICSE)*, pages 617–627, Zurich, June 2012. IEEE.

[123] Koen Yskout, Riccardo Scandariato, and Wouter Joosen. Do security patterns really help designers? In *Proceedings of the 37th International Conference on Software Engineering - Volume 1*, ICSE '15, pages 292–302, Florence, Italy, May 2015. IEEE Press.

A Included Papers

Acar et al. [2], Acar et al. [1], Acar et al. [4], Assal et al. [6], Baig et al. [11], Braz et al. [16], Cohney et al. [24], Danilova et al. [28], Danilova et al. [26], Danilova et al. [27], Derr et al. [29], Dietrich et al. [31], Fahl et al. [37], Fischer et al. [41], Fischer et al. [40], Fulton et al. [43], Gerlitz et al. [44], Gorski et al. [47], Gorski et al. [46], Hänsch et al. [54], Hallett et al. [52], Haney et al. [53], Krombholz et al. [66], Krombholz et al. [65], van der Linden et al. [112], Mhaidli et al. [71], Nadi et al. [73], Naiakshina et al. [77], Naiakshina et al. [76], Naiakshina et al. [75], Naiakshina et al. [74], Nguyen et al. [78], Nosco et al. [80], Oliveira et al. [82], Oltrogge et al. [83], Palombo et al. [85], Pashchenko et al. [86], Plöger et al. [88], Roth et al. [94], Ruef et al. [95], Sheth et al. [100], Tahaei et al. [102], Tiefenau et al. [106], Tiefenau et al. [105], Tuladhar et al. [110], Votipka et al. [116], Votipka et al. [115], Votipka et al. [114], Weir et al. [118], Weir et al. [117], Yakdan et al. [120], Yong Wong et al. [121], Yskout et al. [122], Yskout et al. [123]

B Database Structure

Figure 5 shows a simplified ERD of the database.

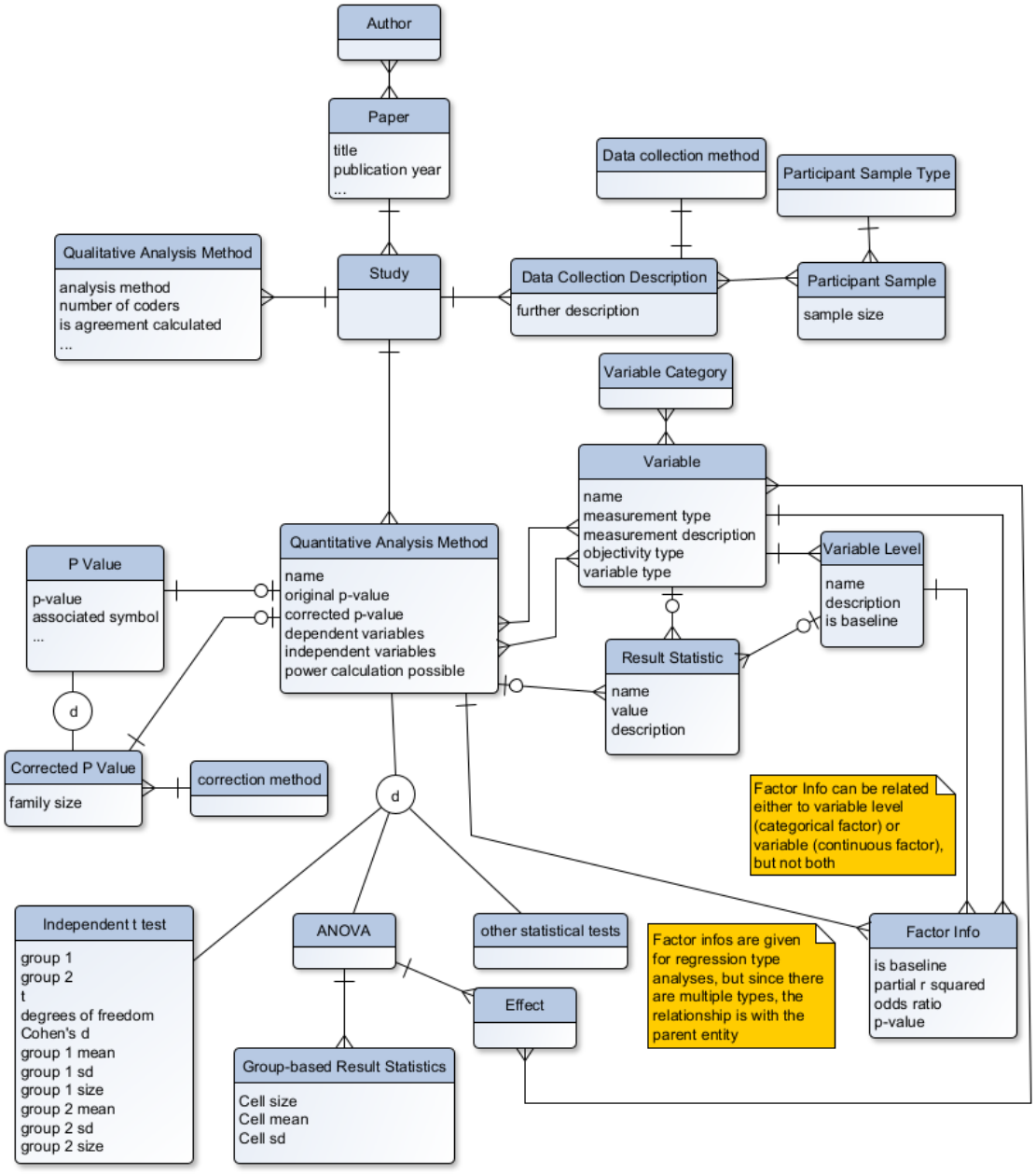


Figure 5: Simplified ERD of the database, not all attributes and entities are depicted.

C A Guide to Power Analysis for Hypothesis Tests with One Categorical Independent Variable with Two Groups

Four parameters are relevant to power analysis: Power, the significance criterion (i.e. the α error level), the reliability of the sample results or sensitivity of the test, and the effect size [22]. The power of a statistical test is the probability of the test correctly rejecting the null hypothesis, i.e. that a statistical test yields a significant result, when the alternative hypothesis is true [35]. Power can also be represented as $1 - \beta$, wherein β is the Type II error, i.e. wrongly rejecting the null hypothesis. This means that if a test has a statistical power of 0.8, as is an often used, acceptable value [22, 30], an actual effect will be detected 80% of the time. The significance criterion or significance level represents the threshold of maximum accepted probability of making a Type I error, i.e. wrongly assuming the alternative hypothesis, detecting an effect, when there actually is none [22]. Using the widely accepted threshold of 0.05 for statistical significance means that only in 5% of cases, an effect is detected in the sample, even though in the population, it does not exist. Reliability refers to how well a sample estimate represents the corresponding population parameter [22]. Reliability is influenced by different factors, depending on the type of estimated parameter, such as the quality of the measurement instrument, and controlling sources of variance in the data, which might distract from the effect you are trying to measure [35]. The largest and invariably present influencing factor, however, is sample size [22], such that larger samples produce more consistent and reliable estimates than smaller ones. Finally, the effect size measures the amount of impact of an independent variable on dependent variables, rather than only judging the presence or absence of an effect [35]. There are generally two types of effect sizes: Non-standardized, or simple effect sizes, which represent the size of effect in the units of the outcome variable, and standardized effect sizes which represent the effect size relative to the variability in the sample or population [10]. When comparing two means, e.g. with a t-test, the difference in mean completion time between two different interface variants represents a simple effect size, measured in units of time, e.g. minutes, while a standardized effect size for this scenario, such as Cohen's d, takes into account the standard deviation in the two groups. Standardized effect sizes are commonly classified as either belonging to the d-family, such as Cohen's d in the example above, or as belonging to the r-family, such as the correlation coefficient Pearson's r [92].

These four parameters are interdependent, such that when three of them are available, it is possible to calculate the fourth. Such calculations are referred to as power analysis. In general, there are four different kinds of power analysis, each used to determine one of the parameters from the other three, although it is also possible to determine both α and power if a ratio for α and β is given together with the other two parameters - this is termed compromise power analysis [38]. The other four flavors are summarized, e.g. by Cohen [22] in Chapter 1.5.

The tutorials on the companion website[5] provide an overview of the data necessary to conduct power analysis for basic hypothesis tests, where to find this data in our database, and how to use it to conduct power analysis using G*Power or R.

[5] powerdb.info/

GuardLens: Supporting Safer Online Browsing for People with Visual Impairments

Smirity Kaushik[1], Natã M. Barbosa[1], Yaman Yu[1], Tanusree Sharma[1], Zachary Kilhoffer[1], JooYoung Seo[1], Sauvik Das[2], Yang Wang[1]
[1]University of Illinois at Urbana-Champaign [2]Carnegie Mellon University
{smirity2, natamb2, yamanyu2, tsharma6, dzk2, jseo1005, yvw}@illinois.edu, {sauvik}@cmu.edu

Abstract

Visual cues play a key role in how users assess the privacy/security of a website, but often remain inaccessible to people with visual impairments (PVIs), disproportionately exposing them to privacy and security risks. We employed an iterative, user-centered design process with 25 PVIs to design and evaluate GuardLens, a browser extension that improves the accessibility of privacy/security cues and helps PVIs assess a website's legitimacy (i.e., if it is a spoof/phish). We started with a formative study to understand what privacy/security cues PVIs find helpful, and then improved GuardLens based on the results. Next, we further refined Guardlens based on a pilot study, and lastly, conducted our main study to evaluate GuardLens' efficacy. The results suggest that GuardLens, by extracting and listing pertinent privacy/security cues in one place for faster and easier access, helps PVIs quickly and accurately determine if websites are legitimate or spoofs. PVIs found cues such as domain age, search result ranking, and the presence/absence of HTTPS encryption especially helpful. We conclude with design implications for tools to support PVIs with safe web browsing.

1 Introduction

Visual cues play a key role in how users assess the privacy/security posture of a website [17] but are often inaccessible to people with visual impairments (PVIs) [32, 33]. In turn, PVIs are disproportionately susceptible to a broad range of security risks, such as phishing threats [12, 17] and challenges with web authentication [18, 27] intertwined with privacy risks, such as shoulder surfing [4] and accidentally sharing personal information [6, 7, 42]. Prior research [1, 41] suggests that it is often difficult for PVIs to assess a website's credibility due to the poor accessibility of privacy/security cues, such as whether a website is HTTPS-enabled.

Our work explores ways to make website privacy/security cues more accessible to PVIs. We followed an iterative, user-centered design approach in designing and evaluating a browser extension, GuardLens, that collects and presents key privacy/security cues for a website, so users do not need to perform these checks manually. Based in part on prior work [1, 10, 32] as well as a formative study and pilot study that explored how PVIs assess the privacy/security posture of a website, GuardLens highlights key privacy/security cues. For instance, some security cues from GuardLens, like domain age registration and search result ranking, are relevant to phishing detection, while cues like HTTPS encryption and website owner are valuable general security cues. Guardlens also provides privacy cues, such as whether website images contain Not Safe For Work (NSFW) content to mitigate shoulder surfing and maintain social norms. Together, the privacy/security cues from GuardLens highlight many privacy and security-related threats to the PVIs online.

We aim to answer two main research questions:

- RQ1: How does GuardLens make privacy/security cues of a website more accessible to PVIs?
- RQ2: How does GuardLens help PVIs assess whether a website is legitimate or a spoof?

We conducted our research iteratively in three stages with 25 PVIs: a formative study (n=5), pilot study (n=3), and main study (n=19). The main study is an experiment (lab-based interview study) that builds on the field study and the pilots to directly answers the research questions. In the main study, participants evaluated the accessibility of privacy/security cues and website legitimacy with and without GuardLens.

Results. Our work has yielded novel and significant results.

First, in one easily accessible location, GuardLens presents important privacy/security cues about a website: e.g., whether it is HTTPS-enabled, its domain age, search result ranking. Without GuardLens, PVIs often miss these cues due to inaccessibility or inconvenience.

Second, GuardLens helps PVIs to determine the legitimacy of websites (spoof or not). Participants found privacy/security cues from GuardLens helpful in correctly determining that spoofs were spoofs and that legitimate, popular sites were

not spoofs. However, it also increased their concerns with unpopular sites that were not spoofed. We reflect on the ways future designs can improve the interpretability of these cues.

Third, we observed novel strategies participants used to assess a website's legitimacy without GuardLens. Strategies included externally verifying that a website's URL is highly ranked in a Google search of its title, checking for links related to copyright information and privacy policy in the footer, and reading URLs character-by-character with a screen-reader.

Contributions. This work makes three main contributions: we (1) designed a new tool, GuardLens, to make the **privacy/security cues of a website more accessible** to PVIs; (2) identified privacy/security cues that participants found useful to determine website's legitimacy while using GuardLens and observed **novel strategies** used by our PVI participants to assess website legitimacy while web browsing; and, (3) offer **recommendations to further improve the accessibility of privacy/security cues** for PVIs.

2 Related Work

Prior literature has studied the privacy and security concerns of PVIs extensively [2, 3, 5, 7, 9, 18, 25]. Researchers have highlighted various privacy concerns for PVIs, such as shoulder surfing [4] and accidentally sharing personal information [6, 7, 42] while browsing websites online. The privacy risks often intertwine with various security risks to PVIs, such as email and website phishing threats [12,50], challenges with web authentication [18,27], and the inaccessibility of security cues [10,32]. For instance, Barbosa et al. [10] highlights that although many websites offer visual cues to facilitate access to features, e.g., log-in, such visual shortcuts are not accessible to PVIs. Similarly, Napoli et al. [32,33] found usability and accessibility issues with online resources, e.g., insufficient web browser security indicators and poor accessibility of password managers. It results in poor access to privacy and security-related information online for PVIs, making them vulnerable to various privacy and security risks, such as unauthorized access to personal information and phishing threats.

2.1 Phishing Threats to PVIs

Phishing is a common problem. The Anti-Phishing Working Group (APWG) [8] detected 266,387 phishing websites in 2019, the highest number since 2016. A large body of work has explored *phishing websites* [15, 23, 28, 30, 34, 37, 38, 45, 47, 51, 52]. Xiang et al. [49] identified two major criteria of a phishing site: a) visual similarity to a legitimate site and b) at least one login form for users to input their credentials. Dhamija et al. [17] found that some phishing sites fooled 90% of participants, and existing anti-phishing browsing cues were ineffective. For instance, studies [19, 39] highlight that phishing websites are increasingly using HTTPS. Consequently, checking whether a website is HTTPS protected is no longer effective against phishing. A study on spear phishing emails found that older adults were more vulnerable to phishing attacks than younger adults [36].

Few studies have explored phishing threats specific to PVIs. Blythe et al. [12] investigated the response of blind users to phishing emails and found they used robust strategies for identifying phish based on a careful reading of emails. However, Abdolrahmani et al. [1] found that it is more challenging for PVIs to assess the credibility of phishing sites because of the inaccessibility of security indicators. Sonowal et al. [41] found similar accessibility issues while evaluating browser extensions designed to protect PVIs against phishing websites.

2.2 Website Privacy/Security Cues

Researchers have examined the effectiveness of privacy/security cues and often found them lacking [17, 29, 43]. Dhamija et al. [17] found that 23% of the participants did not look at browser-based cues such as the address bar, status bar, and security indicators, leading to incorrectly assuming phishing websites safe 40% of the time. Other studies have focused on accessibility issues of privacy/security cues for PVIs. Sonowal et al. [41] found a range of accessibility issues for PVIs, such as color-based privacy/security indications, missing instructions, and lack of shortcut keys. Napoli et al. [32] found that passive browser chrome indicators did not help PVIs browse websites securely because they can only see a small portion of a website when using a screen magnifier. The small field of view is more likely to focus on page content than other areas of the browser. Instead, to comprehend the page as a whole, they skimmed pages while completing tasks and skipped over large portions of content to find relevant information from a website. It is insufficient to provide alternative text to describe security cues like lock icons and SSL certificates because users may not actively seek out this information. As a result, the security information can potentially go unnoticed by users.

2.3 How PVIs Assess Site Credibility?

Researchers [24, 31, 32] have observed that blind and sighted users absorb information differently. Sighted users comprehend information from whole to part. They see the whole picture simultaneously and understand the different visual encodings in relation to each other (e.g., identifying a website as a shopping site upon visiting). In contrast, PVIs put together each piece of information to make sense of the picture as a whole (e.g., scrolling through the webpage to explore what the website is about). They often rely on text and use fast tab/scroll down the webpage as an exploration tactic to find relevant information. In this process, screen-reader users skip over large portions of the content to alleviate heavy cognitive loads associated with browsing websites audibly. However, studies suggest [1, 32] that this habit could increase the like-

lihood of missing vital privacy/security-related information, making it challenging for PVIs to assess webpage credibility [1].

Overall, prior work [10, 25, 41, 46] suggests PVIs are often exposed to privacy and security risks online, including phishing, due to poor accessibility of websites and insufficient privacy/security indicators. These insights informed GuardLens' design.

3 GuardLens System Design

We developed GuardLens with two design goals: (1) to *provide quick access to privacy/security information,* such as a website's domain name, and whether it is HTTPS enabled; and (2) *to equip users with information needed to protect them against privacy/security risks* such as phishing attacks. These goals correspond to helping PVIs overcome the *awareness* and *ability* barriers that can hinder users' acceptance of expert-recommended best practices for security and privacy [16]. Details of the design considerations are in the appendix 8.

3.1 System Overview

GuardLens JS was developed in ES6, compiled with BabelJS, and is executable and tested on Chrome, Firefox, Safari, Edge, and IE10+. We incorporated remote backend development used by the browser extension in response to any requests, local storage to handle data requested from API services, general helpers and algorithms to run required design features, and messages/prompts for users to make informed decisions. Requests to the backend were made over HTTPS, and the endpoints required no user data. For example, the endpoint to return TLS certificate information only requires a URL request parameter. The backend was hosted securely in our university servers with restricted access to our research team. (see Figure 2 in the Appendix). The workflow contains client requests sent to different services and a synchronous process of the data in server endpoint to present the results in the UI. We have open-sourced GuardLens [1]. GuardLens is an unlisted browser extension; only recruited participants received a download link.

The GuardLens web browser extension interacts with backend API endpoints, and the app engine creates queries from user requests (see Figure 2 in the Appendix). The app engine backend receives requests from a script embedded into the browser extension. These requests are sent automatically from the browser extension to the system's backend, where the system has endpoints to each of the cues supported by the browser extension. TensorFlow JS[2], Universal Sentence Encoder [14], and NSFW JS[3] are some of the notable helpers/algorithms used in the backend to build the privacy/security information features.

GuardLens also uses local storage to save users' preferences if they choose to opt out of (1) seeing a GuardLens pop-up for a particular website, or (2) seeing a particular type of information block for all sites in the future.

3.2 Interface Details

Guided by our design goals, we implemented GuardLens as a technology probe [26] that gives users easy access to privacy/security-related cues about a website upon request. GuardLens is a browser extension that consists of a collection of cues meant to surface pertinent privacy and security information about the website that one is currently browsing.

After installing GuardLens, participants read and reviewed the privacy policy for our study, and how data will be used for this research. Participants then chose whether they would consent to start using Guardlens or wish to uninstall it (see details in Figure 1 in Appendix). After users consented on this disclosure interface, they were prompted with a user input field to provide a participant ID. We used "Screen A" for the consent interface and the prompt message.

Once a participant entered their ID and clicked "OK", the GuardLens main interface ("Screen B" in Figure 1 in Appendix) appeared, displaying the privacy/security information of the website presently in focus in the form of information blocks (see "Screen B" in Figure 1 in Appendix). Each information block consists of an expandable drop-down with a ***"Tool Tip"*** and ***Actionable Suggestions***. Below we discuss each information block in the order presented in the GuardLens interface. We chose and ordered these seven S&P cues based on prior work [20, 32, 40] and findings from both our formative and pilot studies.

HTTPS Encryption: This information block highlights whether or not a site uses HTTPS. Prior work [32] suggests that the HTTPS lock icon and/or SSL certificates are often inaccessible to PVIs. To improve the accessibility of security information, the backend system of GuardLens parses TLS certification when a user visits the site. Users can find two additional messages by clicking the expandable drop-down: a) tool tip: *"Based on the actual information from the website's security certificate"*, and b) actionable suggestion: *"What you can do: You may choose not to send your information to this website such as payment or personal information"* if the site lacks HTTPS encryption.

Website Owner Identity: This information block identifies the entity that owns the website. We included this cue for reasons similar to adding the HTTPS information. Additionally, browsers share this information when displaying certificate information, and participants in the formative study found it useful. Clicking into the expandable drop-down, users can find two additional messages: a) tooltip: *"Based on this website's security certificate"*, and b) actionable suggestion:

[1] GuardLens source code: https://github.com/guardlens22/GuardLens
[2] https://www.tensorflow.org/js
[3] https://nsfwjs.com/

"What you can do: You may be more cautious about sending your information to this website not knowing who owns it.". The backend app engine parses the site's TLS certification to extract this information.

Domain Name: This information block states the domain name of the website word for word. We included this cue because some phishing URLs try to confuse users about the domain name, e.g. bestbuy.greatshops.com. Clicking into the expandable drop-down, user find two additional messages: a) tool tip:*"Based on this page's address"*, and b) actionable suggestion: *"What you can do: You may leave this website if it is not the intended website you wanted to access"*. The backend app engine parses the site's URL to extract this information.

Search Result Ranking: This information block presents a website's rank in Google search results. In the formative study and the pilots, participants evaluated the legitimacy of a site by manually checking if the site's domain appears in the top 5 of a Google search of its title, suggesting a need for the cue. Clicking into the expandable drop-down, user can find two additional messages: a) tool tip:*"Based on website title, search results are from Google search"*, and b) actionable suggestion: *"What you can do: If the website does not appear in the top 5 search result it is more likely to be a phish. If you are uncertain, do not enter any personal information."* The backend app engine submits a search with the website title as the query term via Google search APIs and determines whether the site is in the top 5 of the returned results. To the best of our knowledge this cue works with most top websites.

Domain Registration and Age: This information block shows *"The website domain was registered 27 years ago."* (see Screen B (Figure 1). We included this cue based on participants' suggestions from the formative study and the pilots. Clicking into the expandable drop-down, users can find two additional messages: a) tool tip:*"Based on website domain registration"*, and b) actionable suggestion: *"What you can do: Research suggests that younger websites are more likely to be phish. In particular, most phishing sites are less than 2 years old."* We added this actionable suggestion for domain age based on prior phishing studies [22,34,35,44]. The backend app engine parses the site's domain registration from the Prompt API [4] (Whois Lookup API that provides registration details) to extract this information.

External Links: This information block indicates how many external links point out of the website. We included this cue for two reasons. First, phishing sites often reuse the HTML code of the legitimate site they are attempting to spoof, change the part that launches the phishing attacks (e.g., login) and leave the rest intact, which means they often have many links pointing to the original site. Second, deceptive sites with click bait often have many external links [48]. Clicking into the expandable drop-down, user can find two additional messages: a) tool tip:*"Based on the destination address of all links"*, and b) actionable suggestion: *"What you can do: If this number is high, you may want to pay close attention to links before clicking them. You also leave the website if you think it is deceptive or masquerading as a real website, such as a fake website with links that point to the real website."* The backend app engine parses the site's HTML code to identify and count the external links to derive this information.

Image Description: This information block shares whether images on the screen show Not Safe For Work (NSFW) content. We included this cue because the unexpected inclusion of NSFW content can be a signal to help PVIs assess if they are browsing the website they intended to. Moreover, particularly in cases where the PVI may be near bystanders, they can use this information to assess whether or not they should leave the website up in keeping with the social norms of their situation. Clicking into the expandable drop-down, the user can find two additional messages: a) tool tip:*"Based on an automated standard detection of indecent or inappropriate images on the screen, which suggest images show content that may not be safe for work."* as a tool tip, and b) actionable suggestion: *"What you can do: You may want to leave this page if you are not comfortable with potential bystanders seeing your screen."* The backend uses existing trained machine learning models for detecting objects in images and image safety features (e.g., NSFW JS).

4 Methodology

We followed an iterative user-centered design process with a series of three studies: initial formative study, pilot of the main study, and the main study. This research is IRB approved. Our interdisciplinary team has expertise in privacy/security, human-computer interaction, and accessibility. One team member self-identifies as a person who is blind.

4.1 Main Study

We conducted lab-based interview experiment to explore the two main research questions stated in Section 1. These research questions were informed by the results from the formative study and the subsequent pilots. In the formative study, which included five participants, we deployed GuardLens as a technology probe [26] to field-test usefulness of privacy/security cues users while browsing websites. We then improved GuardLens design based on the results to better support PVIs needs. Next, we **piloted** the new design with three participants and made further improvements. Finally, our **main study** included 19 participants. Details of the formative study and the pilot study are included in the appendix 8.

4.1.1 Study Design

Due to the pandemic, we conducted the study remotely using Zoom. The one-hour session began with the study tasks

[4] https://www.crunchbase.com/organization/prompt-api

embedded within the interview questionnaire, followed by an exit interview. Participants received a $30 USD gift certificate upon completion of the session. The study tasks followed a within-subject design, where all participants browsed six websites. These websites were selected from three categories: popular, unpopular, and spoof across seven domains: finance, e-commerce, accessibility, news/media, education, healthcare, and productivity. ***Popular sites*** were chosen from sites in the top 1,000 Alexa ranking[5]. ***Unpopular sites*** were chosen from sites ranked 5,001+ in the Alexa ranking. The popular and unpopular sites are not spoofs. For ***spoof sites***, we developed spoofs of two popular sites for our target user population: amazon.com (Amazon) and nfb.org (National Federation of the Blind). We created these spoofs to be visually similar to their legitimate counterparts, similar to prior studies [33, 49]. The domain names of the spoof sites sounded identical to the original sites when read out aloud by a screen reader but were spelled differently: i.e. amaZaunn.com vs. amazon.com. Also, these spoof sites were safe to browse. Feedback from the formative study suggested GuardLens' usefulness depends on the popularity of and familiarity with the site. We thus explored these factors in the main study by having participants visit six websites that varied in familiarity and popularity, simulating real-world browsing. It helped us to test GuardLens' effectiveness at assessing site security, privacy features, and legitimacy across popular (often familiar), unpopular (often unfamiliar), and spoof (of popular) sites. Table 4 (Appendix) lists all the websites used in the main study.

For the study tasks, we emailed participants links to the websites we chose. We followed a scenario-based approach, commonly used in the prior work on phishing [17]. Our scenario stated, *'Imagine that you receive an email message that asks you to click on one of the following six website links. Imagine that you decide to click on the link to see if it is a legitimate website or a "spoof" (a fraudulent copy of that website). Please browse three websites using the GuardLens tool and the other three without the tool.'* We randomly selected two popular and two unpopular websites from a pool of four popular and four unpopular sites (see Appendix Table 4). The same two spoof sites were presented to all participants. Each participant browsed three sites (one popular, one unpopular, and one spoof) with GuardLens and another three sites without GuardLens without knowing the conditions (popular, unpopular, spoof sites). Note that, we counterbalanced the order of presentation of websites using Guardlens and without it. Some participants were first presented with GuardLens, followed by browsing websites without it and vice versa.

After browsing each website, participants were asked five 5-point Likert scale questions and three open-ended questions [6]. The Likert scale questions asked participants to rate legitimacy, familiarity, accessibility, ease of assessing privacy and security of the website, and whether they would recommend the site to their friends. They were also asked to provide reasoning for each rating. We also asked participants an open-ended question about the strategy they used to detect the privacy/security features of the website. If a participant read the URL of the website character by character, we asked open-ended questions about what prompted them, and how often they do so in daily life.

After participants completed browsing the six sites and answering the questions, which took about 45 minutes, we ended the study with a 15-minute exit interview. In the exit interview, we asked participants open-ended questions about their experiences of browsing the sites with and without GuardLens. Figure 5 (Appendix) illustrates the main study design.

4.2 Participants.

We recruited participants through the National Federation of Blind (NFB) mailing list and Reddit (r/Blind). Prospective participants took a screening survey with basic information on age group, occupation, self-reported visual abilities, and their regularly used email services, browsers, and screen readers. Eligible participants must (1) self-identify with visual impairments and (2) regularly use screen readers and the Chrome browser. The goal was to ensure that participants were familiar with the technical environment we provided. Then we identified 19 eligible participants (nine female, 10 male) to participate in our interview session (see appendix Table 3). 15 participants self-described as individuals who are blind and the other four self-described as individuals with low vision. All 19 participants used screen readers. Only P17 did the formative study and no participants did the pilot study.

We provided participants an online consent form within the screening survey, informing about our study procedure and data protection policy. We informed participants that this study was designed to improve the accessibility of privacy/security of browsing websites online.

4.3 Ethics

Our study was approved by our IRB. Prior to each of the three studies, participants signed a consent form, including an agreement to audio/video record. At the start of each session, we re-confirmed their consent and communicated our pseudonymization procedure. We also reminded them their participation was entirely voluntary.

4.4 Data Collection and Analysis

Upon receiving participant consent, we asked them to share their screen and began recording. We also took notes during the study. Our analysis was driven by our main research questions. To answer our questions on the ease of accessing

[5] Alexa Internet was a web traffic analysis company, owned by Amazon. It was discontinued on May 1, 2022. https://www.alexa.com/

[6] GuardLens study questions: https://github.com/guardlens22/GuardLens

privacy/security cues on a website, assessing a website's legitimacy, and security strategies participants employed, we first qualitatively analyzed participants' responses using thematic analysis [13]. Two co-authors (coders) manually and independently generated initial codes that capture meanings of the same subset of our interview data at a fine-grained level (usually at the sentence level). Then, the two coders discussed, and converged their codes into a code book of 50 unique codes ranging from easy access to GuardLens, trusted website footer links, and familiarity with site. We calculated the inter-coder reliability is 0.88 (Cohen's Kappa), which is considered good [21]. Next, the two coders used the agreed-upon code book [7] to code the rest of the responses. We followed an open coding method to explore how participants used GuardLens and why they found it helpful or not. We added new codes to the code-book when existing codes could not capture the data, until the code saturation was achieved. We then grouped all codes into higher-level themes, such as tool support, legitimacy assessment, and website content familiarity.

We next employed quantitative methods to assess if use of GuardLens resulted in statistically significant differences in: (1) participants' perceptions about the accessibility of privacy/security cues on a website; and, (2) participants' ability to differentiate between legitimate and spoofed websites. We also explored how independent factors — such as the accessibility of a website and participants' familiarity with the website — impacted users' ratings for assessing a website's privacy/security and legitimacy. To do so, we employed a mixed-effects regression analysis (R lme4 [11] package): we included participants' familiarity and perceived accessibility of a website as covariates, participants' use (or not) of GuardLens as the independent variable, and included a random-intercepts term for participant IDs since each participant browsed and rated multiple sites.

5 Results

We first examine participants' perceived ease of accessing privacy/security cues with or without using GuardLens for three types of websites: spoof, popular (legitimate), and unpopular (legitimate) (RQ1). Next, we evaluate whether participants correctly determine the website's legitimacy (i.e., spoof or not) with or without using GuardLens for each type of website (RQ2). We hypothesized that GuardLens should make privacy/security cues more accessible and help PVIs more easily assess website legitimacy.

5.1 Ease of Accessing Privacy/Security Cues

We asked participants to rate and provide reasoning for the ease of accessing the privacy/security cues of a website on a 5-point Likert scale (the "ease rating"). Ratings 4 and above mean participants found it easy to access the privacy/security cues; ratings 2 and below indicate that participants found it difficult, and a rating of 3 indicates neutrality. Figure 3 in appendix 8 shows the ratings for different types of websites with or without GuardLens. Table 1 in appendix 8 summarizes the most accessible privacy/security cues participants used with or without GuardLens.

5.1.1 Spoof Sites

We hypothesized that PVIs would access privacy/security cues on spoof websites more easily with GuardLens than without. Our results confirm the hypothesis.

Each participant visited a spoof of two sites, Amazon and the National Federation for Blind (NFB), which are well-known to our target user populations. If participants were asked to browse the spoof NFB site (eneffbee.org) using GuardLens, then they would browse the spoof Amazon (amazaunn.com) without using GuardLens and vice versa.

We used linear mixed-effect regression analysis to determine how GuardLens impacts participants' perceived ease of accessing privacy/security cues. The ease rating was the dependent variable, while using GuardLens or not was the independent variable. The familiarity rating and the accessibility rating of the site from the specific participant were covariates. We also included a random intercept term for each participant ID to account for repeated observations. The R lme4 model is: $ease = tool + familiarity + accessibility + (1|pid)$

Finally, we estimated the statistical significance (p-values) of the fixed effects with the R car::anova function (type III Wald Chi Square test). The evidence suggests that GuardLens made privacy/security assessments easier for PVIs as they browsed spoof websites. Participants gave significantly higher ease ratings when browsing spoof sites with GuardLens than without (estimate coefficient = 0.9152, $p<0.05^*$). Below, we present qualitative results providing additional context for why, and distill our findings into a key takeaway.

Without GuardLens, about 47% of participants gave a rating of 4 or above, while 53% gave a rating of 3 or below for ease of accessing privacy/security cues on spoof sites. It suggested that participants found it difficult to assess the privacy/security of spoof sites without GuardLens' cues.

Six participants (33%) checked the website's URL character by character using a screen reader, which helped them determine the site was a spoof. While three out of these six participants habitually checked for URLs character by character, the other three were primed by the URL's odd pronunciation.

Some participants checked a combination of specific privacy/security cues. For instance, those (16%) who searched the site for layout and footer information (e.g., contact us, privacy links, and copyright information) also checked for

[7] GuardLens study codebook: https://github.com/guardlens22/GuardLens

HTTPS. For instance, P11 gave an ease rating of 4 for the spoof NFB site because, *"It's a familiar website. I recognize the link, and there is https on the top"*. Note that participants gave an ease and legitimacy ratings independently; in this case, though the NFB site was spoofed, the participant still believed it was easy to access privacy/security cues without GuardLens, and ultimately made an incorrect determination.

With GuardLens, the majority (74%) of participants rated ease of accessing the privacy/security cues on spoof websites 4 and above. Participants stated GuardLens cues about a website's domain age and (lack of) appearance in the top five Google results raised suspicion, prompting them to manually check the URL character-by-character with their screen reader. For example, P8 rated ease of accessing privacy/security cues a 5 when visiting the spoof NFB site (eneffbee.org) based on the cues from GuardLens because, *"It was easy. It (website) was registered 10 months ago, 527 links go to other websites, and I spelled the URL—that's not them."*

Unlike P8, P13 ignored the GuardLens cues on the spoof amazaunn.com site. She checked all the cues, then stated *"I can't understand why GuardLens stated domain age as 11 months."*, as this information contradicted her expectation about Amazon's age. P13 assumed that Amazon's security certificate was renewed 11 months ago, then ignored Guardlens' domain age warning and assessed the site as credible based on the website footer links. This finding suggests that when GuardLens cues contrast with user expectations, some users may doubt the cue itself. We articulate relevant design implications for GuardLens in the discussion.

Observation 1: For spoof sites, GuardLens cues prompted many PVIs to check the URL character by character, making it significantly easier for them to assess the privacy/security of these websites.

5.1.2 Popular Sites

We hypothesized that people with visual impairments would rate ease of accessing the privacy/security cues on popular websites to be higher when using GuardLens than when not. Our results support this hypothesis. When using GuardLens, participants rated the ease of accessing privacy/security cues on popular websites significantly higher (estimate coefficient = 1.208, $p < 0.0005^{***}$). We highlight participants' reasoning for preferring GuardLens and provide a conclusion in observation 2.

Without GuardLens, approximately 47% of participants rated ease of accessing privacy/security cues of a popular website 4 and above. They often relied on checking the HTTPS encryption in the URL and the website footer information, such as the presence of copyright and privacy links. Those who gave ratings of 3 and below (53%) were unsure how to check a website's privacy/security cues.

With GuardLens, approximately 95% of participants rated ease of accessing privacy/security cues 4 and above for popular sites. Most relied on GuardLens because the tool consolidated website's security-related information in one place. For instance, P19 said*"Everything I needed to know about the website was in one place. I didn't have to look at all other places. It was a lot easier."* Participants further reported they found GuardLens' security information accurate and trustworthy. P7 said they browsed the website footer and found the *"Copyright info matched with GuardLens domain age."* Some of the security cues from GuardLens that participants found particularly helpful were domain age and HTTPS encryption information.

Observation 2: For popular (legitimate) sites, GuardLens significantly eases PVIs' access to a site's privacy/security cues by consolidating them in one place.

5.1.3 Unpopular Sites

We hypothesized that people with visual impairments would rate ease of accessing the privacy/security cues on unpopular websites to be higher when using GuardLens than when not. We did not observe strong evidence to support this hypothesis. While the descriptive statistics show that people gave higher ratings using GuardLens, using GuardLens was not a significant factor in the mixed-effect regression model ($p>0.05$). We further explore why by evaluating participants' reasoning and provide a conclusion in observation 3.

Without GuardLens, 36% of our participants gave ratings of 4 and above for ease of accessing the privacy/security cues of a website. Participants in this rating group often checked for three cues: HTTPS encryption in the URL; the presence of a privacy policy link in the website footer; and the general readability, accessibility, and layout of the website. For example, while browsing a productivity site (openoffice.org), P18 reasoned that it was easy for him to assess the privacy/security of the site because *"(The site was) built like other ones, and there's privacy policy link."* He was not familiar with the site so he browsed it thoroughly and found it accessible, similar to the other websites he often visits.

Some participants provided unique reasoning for their rating. While browsing a money-transferring site (zapsend.com), P4 reasoned that the website appeared in the top 5 Google search results, so it was easy to assess its privacy/security. Although he navigated through the website, he did not rely on the features within the site to assess its privacy/security. Rather, he verified whether it was a spoof or not by googling it and then matching the URL of the search result with the website we gave him to browse. Another participant (P11) visiting a shopping site (zolucky.com) accessed the website's SSL certificates by clicking on the lock icon near the address bar to check its domain registration date. Since he found that the website was registered and the security certificate was valid, he gave the rating 5 for ease of accessing privacy/security of

the site. Interestingly, P16 assessed the privacy/security of the same shopping site based on customer reviews for its products. *"It didn't take me a lot of time to realize there were no customer reviews. The cursor kept moving around."* She also found that the website had poor accessibility features, and the website footer did not include a privacy policy link. She concluded *"Even if it has https, I wouldn't trust it."* While we focused on assessing how well GuardLens helps participants identify phish, participants also assessed other types of threats. For example, here the site was not a spoof, but still seemed untrustworthy to this participant.

21% participants gave a rating of 3, and 47% participants gave a rating of 2 and below because they were unfamiliar with the website and uncertain of their assessment. For instance, P5 rated an unpopular audiobook site 3 *"because I am not familiar with the website. I am not very knowledgeable on website security and domain."* Similarly, P14 and P15 were uncertain because they were unaware of what type of data the sites collected from them. However, while browsing an online learning site from another country, P15 felt skeptical, *"I'm not certain of my assessment, it was much more difficult. I have my own biases because it's in Nigeria. I would be hesitant to buy something from a website in another country."* In the case of a financial money transfer website, P12 mentioned that *"I think with all these websites, it's very hard just by looking at it without entering personal information."* Participants also googled the websites; and checked for layout, content, and accessibility. P9 said, *"the score goes down because I couldn't find a Google result with website link. But the actual website looked legitimate."*

With GuardLens, more participants (42%) gave a rating of 4 and above for ease of accessing privacy/security cues of unpopular sites. It suggests that although we observed mixed results about the effectiveness of GuardLens on unpopular sites, the tool improves accessibility. Participants relied on GuardLens to access privacy/security cues about the website. However, even with GuardLens, they found it tougher to assess the privacy and security of unfamiliar websites. Those who gave ratings 3 (26%) or 2 and below (32%) found the information from GuardLens confusing, especially for unpopular sites hosting illegal content such as audiobook torrents. For example, while browsing an audiobook site (http://audiobookbay.ws/), P2 said *"It was difficult because the info in GuardLens was contradictory. It was in the top 5 search results and had low external links but it also had warnings. It was not clear to me. They might be illegally sharing audiobooks but not really trying to get my information."* According to P2, although GuardLens suggested two positive features for the site, it also gave warnings such as the site lacks HTTPS encryption, and the site has a younger domain, suggesting that it may not be safe. In such cases, even though GuardLens provided access to privacy/security information, it was insufficient. An important note: by "legitimate" websites, we mean sites that are not spoofs — not that the website is "secure" and harm-free. The audiobooks website in this example hosts torrents for audiobooks which is illegal in the US. However, the website is still safe to browse unless the user downloads anything from it. In that case, maybe they could download some potentially malicious files.

Observation 3: GuardLens privacy/security cues for unpopular (legitimate) sites are less helpful. Lack of familiarity with a site, and sometimes mixed (positive and negative) cues, seem to complicate user assessments.

5.2 RQ2: Assessing Website Legitimacy

We asked participants to rate the legitimacy of the websites on a 5-point Likert scale, where a high rating (>3) means that the user thinks the website is not a spoof or a phish. We also asked about their reasoning for the rating, and the security strategies used to assess legitimacy across the three website types (spoof, popular, and unpopular). We used the same spoof websites described in Section 5.1.1 for assessing site legitimacy. Figure 4 in appendix 8 shows ratings for different types of websites with and without GuardLens. Table 2 in appendix 8 summarizes the most popular strategies participants used to assess website legitimacy with and without GuardLens.

5.2.1 Spoof Sites

We hypothesized that PVIs would rate the legitimacy of spoof websites lower with GuardLens than without. Our results confirms this hypothesis. We performed a linear mixed-effect regression to determine Guardlens' impact on the perceived legitimacy of a site. The R model is:
$legitimacy = tool + familiarity + accessibility + (1|pid)$

We estimated the p-values of the fixed effects using the car::anova function (type III Wald chi-square test). We found statistically significant evidence suggesting that GuardLens impacted participants' legitimacy ratings for spoof websites (estimate coefficient = -0.8279, $p<0.05^{*}$). Participants gave a lower legitimacy rating for spoof sites when they had access to GuardLens than when they did not. We present their reasoning for the rating and provide a conclusion in observation 4.

Without GuardLens, participants tended to ignore the cues of spoof websites and assessed legitimacy based on their familiarity with the website. Only 45% of participants identified the spoof websites. Among these participants, 39% gave a rating of 2 and below and 6% gave rating of 3. The remaining 55% of participants failed to identify the spoof sites and gave legitimacy ratings of 4 and above.

Participants who successfully identified a spoof site without GuardLens often relied on manually reading the URL character by character using a screen reader. Participants also often checked whether the website was HTTPS-enabled. For example, P7 assessed the spoof website they encountered

without GuardLens as illegitimate: *"I don't think this is legitimate. The URL is very suspicious. But the homepage sounds like its clone."* But for the 55% of participants who failed to identify the spoof websites without GuardLens, they all mentioned familiarity with the site as the main reason for their high legitimacy ratings. P3 assessed the spoof NFB site as being legitimate with certainty, *"I am extremely sure the website is legitimate. I have been on the website (before)."*

With GuardLens, participants used cues which are otherwise inaccessible such as domain age of the website. Only 28% of participants failed to identify spoof websites. By contrast, the majority of participants (72%) successfully identified the spoof websites using GuardLens.

When participants used GuardLens, its cues were the most popular security strategy they used in making their legitimacy assessments. The most commonly cited GuardLens cue was the domain age of the website. Indeed, the domain age cue in GuardLens suggested that if the website was less than 2 years old, the site may be more likely to be a phish.

For instance, when visiting the spoof Amazon site, the tool surfaced that the domain age of the website was 9 months. This cue raised suspicion among participants since Amazon has been in the market for over 20 years. Similar observations were made for the spoof NFB site. For instance, P4 gave a low legitimacy rating (2) for the spoof Amazon site, explaining *"I am not sure at all (whether the website is legitimate). Because it seems to be a legitimate site, but GuardLens said it's a website from 10 months ago. So I'll give 2."*

Among participants who used GuardLens but failed to identify the spoof websites, the most common strategy employed was relying on their familiarity with the website content, layout, and accessibility. Even though they may have noticed suspicion-raising privacy/security cues of the spoof websites on GuardLens, they tended to make their assessments relying on familiarity. For example, in explaining why she gave a spoof site a legitimacy rating of 5, P2 said: *"I read the info provided by the tool which indicated that it was secure. I further confirmed by browsing that it is identical to one I browse."*

Observation 4: GuardLens significantly helped participants correctly identify spoof websites by providing privacy/security cues in one place.

5.2.2 Popular Sites

We hypothesized that PVIs would rate the legitimacy of popular websites to be higher when using GuardLens than when not. Our results confirm this hypothesis. We found significant evidence to suggest that GuardLens affected participants' legitimacy ratings for popular websites (estimate coefficient = 0.6405, $p<0.0005^{***}$). Unlike spoof websites, popular sites are most visited and are legitimate websites. Using GuardLens, participants gave higher legitimacy ratings for popular sites. Below we highlight their reasoning for the rating and provide a conclusion in observation 5.

Without GuardLens, 90% of participants gave legitimacy ratings of 4 and above, 10% of participants gave a neutral rating (of 3). The top three security strategies participants used were URL-related strategies (e.g., reading URL character by character using screen reader, checking for HTTPS encryption), browsing content of websites, and relying on familiarity with websites. For popular websites participants browsed daily, they tended to believe that the website was legitimate. Some of the participants (2 out of 19) did not check security cues but made decisions only based on familiarity. P2 and P3 gave high legitimacy ratings to popular websites. The reasons for their decision were, respectively: *"It's the NY times and it also seems consistent with what I know NYT should be."* and *"I visit it a lot (target.com)"*. In addition, other participants *"manually read URL character by character"* or attempted to *"check if the website uses https encryption"*. Since participants used these popular websites in their daily life, they remembered what the website URL should be. Thus, simple strategies such as comparing URLs could help facilitate participant assessment of site legitimacy.

With GuardLens, participants noticed more security cues instead of relying only on their familiarity with websites and checking URLs. 100% of participants gave legitimacy ratings 4 and above and successfully identified popular websites as legitimate. Participants preferred using GuardLens cues as the most popular security strategy to assess website legitimacy. They found three cues most useful: the website's domain age, Google search ranking, and the presence/absence of HTTPS encryption. For instance, P2 assessed a popular website as legitimate because *"GuardLens shows that it is a secure HTTPS website and has been around for 26 years; most phishing sites are not around that long."* Other than website's domain age and search ranking information, P2 also relied on the website's HTTPS encryption information, even though it is not a helpful cue to assess phishing websites.

P3 also noticed more cues, explaining their high legitimacy rating: *"very easy to navigate, headings were readable and in the right spot."* However, familiarity with websites is still a main factor influencing legitimacy perception. P5 explained *"Based on the content of the website and Guardlens information, I feel it is a real site. I don't know how you can copy an entire domain. But the content seemed familiar. I am familiar with NFB, so it is easy for me to recognize the content."* Interestingly, we found familiarity with websites both helped and hindered participants in correctly identifying legitimate websites.

Observation 5: GuardLens significantly helped participants correctly identify the legitimacy of popular sites. Participants leveraged their familiarity with the site, and GuardLens facilitated their assessment by providing cues (e.g. domain age of the site).

5.2.3 Unpopular Sites

We hypothesized that PVIs would rate legitimacy of unpopular (legitimate) websites higher with GuardLens than without. However, our results suggest the opposite. Using the same linear mixed effect model, GuardLens had a significant (negative) impact on participants' perception of unpopular websites' legitimacy (estimate coefficient = -0.6207, $p<0.005^{**}$). This suggests that GuardLens misleadingly increased participants' concern about the sites' legitimacy. Unlike popular websites, some unpopular websites focus less on privacy and security design. GuardLens helped participants identify security issues in unpopular websites, such as a lack of HTTPS encryption. Participants gave low legitimacy ratings based on security issues and unfamiliarity with unpopular sites. While these unpopular sites are not spoofed, their lack of security protection (e.g., HTTPS) is still worth noting to users. Thus, GuardLens can still be useful by presenting cues for multiple threats. Although the only security threat our study assessed was phishing, participants may have given lower legitimacy rating to certain unpopular sites based on poor security properties of those sites in general. We present participants' reasoning in detail below and conclude in observation 6.

Without GuardLens, 36% of participants gave a rating of 4 and above. 32% of participants gave a rating of 3, and 32% 2 and below. Being unfamiliar with these unpopular websites, participants most often used URL-related strategies to determine legitimacy. Since participants are not familiar with the URLs of these unpopular websites, most of them googled the URL. However, some participants did not realize that some of these websites do not use HTTPS. For instance, P14 gave a legitimacy rating of 5 to http://audiobookbay.ws/ and did not check the site for HTTPS. He stated *"I think this website is audiobook service provider."* In addition, P4, P10, P11, and P18 ignored the lack of HTTPS when they browsed unpopular websites without GuardLens.

With GuardLens, 20% of participants rated legitimacy 4 and above, 45% felt neutral (rating 3), and 35% rated 2 and below for unpopular websites that are not spoofs. GuardLens identified and presented some security issues of these websites, which made participants concerned about these sites' legitimacy. For example, GuardLens helped participants notice some unpopular websites not using HTTPS. P1 said *"(I knew) because the tool told me that it was not secure and warning about encryption."*

Observation 6: GuardLens highlighted security issues (e.g., no HTTPS) in some unpopular websites. These (negative) cues made PVIs significantly more concerned about the website's legitimacy. While these unpopular websites are not spoofs, these security issues still pose threats to users and deserve their attention.

6 Discussion

We employed a user-centered design process to design, implement, and evaluate GuardLens: a web browser extension that helps PVIs make informed privacy and security decisions about a website by surfacing a basket of privacy/security cues that would otherwise be inaccessible. Our results reveal the strengths and limitations of the current design and points to a rich area for future research and design.

Our results suggest that GuardLens improves the accessibility of privacy/security cues on websites and helps PVIs make informed decisions about website legitimacy, especially for spoofed and legitimate popular sites. Prior literature [2,32,41] has highlighted the accessibility issues of these cues. PVIs often miss these cues as they try to piece together and make sense of information on the website as a whole [24,31]. Our participants expressed appreciation that GuardLens, through its varied information blocks described in Section 3, provides a bird's eye view of the privacy/security information of a website in one, accessible location.

Prior studies [1,41] explored accessibility challenges faced by PVIs to identify the credibility of websites in general. Our study explores how this population interacts differently with websites to assess their credibility, depending on whether the website is popular, unpopular, or a spoof site. Our study asked participants to browse those three types of websites to mimic their real-world browsing experience.

GuardLens and Spoof Sites. Prior work [49] has identified two major criteria for phishing (spoof) sites: a) visual similarity to the legitimate site and b) at least one login page for users to input credentials. Our study's spoof sites are visually similar to the original sites for Amazon and the National Federation of the Blind. Using GuardLens, a significant majority of participants identified the spoof sites, relying on tool information such as domain age, search result ranking, and the domain name of the website. However, some participants still failed to identify the spoof sites. While they checked the information provided by the tool, they still relied on familiarity with the website's content and layout based on past browsing experiences with original sites. Two participants ignored the red flags about shorter domain age and website not appearing in the top five search results from GuardLens because they thought GuardLens had some glitches. We will revisit this challenge in the design implications section.

GuardLens and Popular Sites. GuardLens was also effective at helping users assess the legitimacy of popular sites. For example, by validating that the site is among the top Google search results for its title and by confirming that the site domain was registered when the user might have expected, participants could confidently recognize the website as legitimate. GuardLens provides an overview of these privacy/security cues in one location.

GuardLens and Unpopular Sites. Unlike the spoof and popular sites, we observed mixed results using GuardLens for

unpopular sites. Multiple participants stated that GuardLens reduced the effort to identify privacy/security information on a website, consolidating this information in one place. However, since our participants were unfamiliar with these sites, strategies such as checking domain age using GuardLens were not helpful in making legitimacy judgments, because participants did not have apriori expectations. Moreover, GuardLens elevates security cues not to be directly pertinent to whether or not a website is a phish, but nevertheless reveal poor security properties. It could cause confusion, as participants might conflate general security with legitimacy. For example, the presence or absence of HTTPS is not always relevant for assessing website phish [1]; yet, websites without HTTPS are less secure, leaving viewers more susceptible to man-in-the-middle attacks. Nevertheless, some participants relied on the HTTPS cue when making legitimacy assessments.

More generally, GuardLens cues correspond to different privacy/security threats without clear distinction. We will revisit this design challenge in the design implications section.

Security Assessment Strategies. Prior literature [33] touches on the security assessment strategies such as fast tab/scroll used by PVIs to determine a website's legitimacy and overall privacy/security posture. Our results confirm those accessibility-based strategies. However, unlike prior study [1], which claimed that PVIs may not rely on HTTPS or SSL/TLS dialogues to assess whether a website is legitimate or fraudulent, our participants considered the presence of HTTPS encryption in URL an important characteristic of a legitimate website. In addition, we also observed some novel strategies. Our participants relied on the website footer links, which included privacy policy, copyright information, 'Contact Us,' and 'About Us,' to determine website's legitimacy.

They also relied on their experience and familiarity with specific popular sites. They would often compare the content of the site they visited during the study with an impression of the site they had based on familiarity.

6.1 Design Implications

Privacy/security cue explanation. Participants found it challenging to interpret some GuardLens cues (e.g., *website's owner identity is unknown*). While GuardLens includes an expandable summary of what a cue means and what a user can do, our participants did not always check or understand those details. Future research should explore alternative ways to present such information: for instance, a chatbot allowing users to directly ask questions about those concepts.

Website accessibility and footer indicators. Screen reader users utilized a website's accessibility and footer information to assess a website's legitimacy. Browsers and security tools similar to GuardLens should consider adding a score to summarize websites' accessibility. An accessibility score could use factors like heading structure, inclusion of image description (alt-txt), and compatibility with various screen-readers such as JAWS, NVDA, or VoiceOver. Similarly, a footer score could highlight the presence of information such as privacy policy, copyright, and contact information.

Structuring privacy/security cues. GuardLens provides mixed signals for unpopular sites. For instance, for an unpopular audiobooks site, GuardLens warned that the site lacks HTTPS encryption and has a younger domain age, suggesting it may not be safe. However, GuardLens also mentioned that the site appeared in the top five search results and had few external links, suggesting the site is safe. Different GuardLens cues tend to correspond to different threats and might sometimes confuse users. Future designs can more explicitly distinguish the underlying threats (e.g., man-in-the-middle attacks, phishing) and structure the cues accordingly.

Providing a blanket privacy/security statement? Some participants desired a simple blanket statement about whether they should visit a site or not. We believe that tools could provide a strong warning for sites that are clearly problematic (e.g., spoof sites). However, as for the long tail of unpopular sites that often have mixed privacy/security cues, providing such a blanket statement is risky because it does not convey the nuance of privacy/security. In those cases, providing detailed but structured (based on underlying threats) cues might be more appropriate.

Engendering user trust with privacy/security tools. Sometimes participants suspected GuardLens has glitches because the cues conflict with expectations. For instance, when Guardlens suggested that the domain age of a spoofed Amazon site was 11 months. P13 nevertheless fell for the spoof because they thought that GuardLens was wrong, mistakenly showing the age of the site's current SSL certificate. Exploring ways to increase users' trust in assessment tools like GuardLens is another design challenge for future work. One strategy could be to more explicitly state where and how the tool creates a security cue (e.g., domain age). Another strategy is the web browser directly incorporating such features rather than having them in a third-party tool.

6.2 Limitations

6.2.1 Limitations of the Current GuardLens Design

Sound Alerts. Participants suggested that GuardLens should have a sound alert when it pops up on the screen with a warning about website. It would nudge users to check the security cues of a website.

Reading Website Domain Names Character by Character. In the current version of GuardLens, the domain name information block states the website domain name as words (e.g., Amazon). Participants must manually read the name by character using a screen-reader (e.g., A-m-a-z-o-n) to verify the spelling of the domain name. Participants suggested that if GuardLens could read out the domain name of the website character by character, it would help PVIs to more easily no-

tice whether they are visiting a phishing website that uses a domain name similar to a legitimate website. Future design could incorporate an option to automatically pronounce the website domain name character by character.

Activating GuardLens. Several participants preferred GuardLens to pop up only when visiting a new site because they are already certain about sites they frequently visit. In the current version, GuardLens does not filter whether the user has previously visited a site. Future design could explore an option where users can define different policies for enacting GuardLens. For instance, GuardLens could ignore a whitelist of sites that a user visited more than twice in the past month.

Catering to Different Levels of Technical Expertise. Participants exhibited multiple levels of technical expertise in the study. While GuardLens provides privacy/security cues for a website, it does not adjust itself for an individual user's technical expertise. Future iterations of GuardLens could be improved to better cater to individual differences in technical expertise, which could be voluntarily provided by a user at the first time of usage by answering a short set of questions.

6.2.2 Limitations of Our User Study

Sample Size. 25 participants finished our study. While it would be desirable to have more participants with different backgrounds, our sample size is on par with the other privacy/security user studies focusing on PVIs [12, 32].

Study Design. Though atypical, we first conducted the formative field study, followed by the summative lab study. In the formative field study, participants used GuardLens as part of their regular browsing experience. The field study strengthened the system's ecological validity and improved its design. The main study yielded many insights, but we could not test GuardLens in real-world context. Participants in the lab-based interview study were aware of being observed and could have been primed to look for privacy/security cues both with and without GuardLens. Nevertheless, the comparison results remain valid. Future work could conduct another summative field study to observe participants' use of GuardLens in situ. We could only test a few websites and website genres to conduct the study within a reasonable duration, especially because these tasks could be taxing for our participants. Future work could explore additional sites, along with GuardLens's usability, factors influencing its adoption/abandonment, and inclusion of other security and privacy features.

7 Conclusion

To address the accessibility barriers that PVIs face in assessing the privacy/security posture of a website, we conducted an iterative, user-centered design process with 25 PVIs. First, we explored what privacy/security cues PVIs find helpful in assessing the legitimacy of websites. Using this knowledge, we designed and implemented GuardLens, a web browser extension that automates and aggregates these cues for PVIs. We then evaluated if and how GuardLens helps PVIs assess the legitimacy of three types of websites, i.e. spoof, popular, and unpopular. We found that while PVIs had difficulty interpreting GuardLens cues for legitimate, unpopular websites with otherwise poor security properties, it effectively increased the accessibility of privacy/security cues, and was helpful for PVIs in assessing the legitimacy of spoof and popular sites.

8 Acknowledgements

We thank our participants for their contributions and sharing their insights. This research was in part supported by the National Science Foundation (NSF) grants #2126314 and #2028387 and #2126058.

References

[1] Ali Abdolrahmani and Ravi Kuber. Should i trust it when i cannot see it? credibility assessment for blind web users. In *Proceedings of the 18th international acm sigaccess conference on computers and accessibility*, 2016.

[2] Tousif Ahmed, Roberto Hoyle, Kay Connelly, David Crandall, and Apu Kapadia. Privacy concerns and behaviors of people with visual impairments. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 2015.

[3] Tousif Ahmed, Roberto Hoyle, Patrick Shaffer, Kay Connelly, David Crandall, and Apu Kapadia. Understanding the physical safety, security, and privacy concerns of people with visual impairments. *IEEE Internet Computing*, 21(3):56–63, 2017.

[4] Tousif Ahmed, Apu Kapadia, Venkatesh Potluri, and Manohar Swaminathan. Up to a limit? privacy concerns of bystanders and their willingness to share additional information with visually impaired users of assistive technologies. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(3):1–27, 2018.

[5] Tousif Ahmed, Patrick Shaffer, Kay Connelly, David Crandall, and Apu Kapadia. Addressing physical safety, security, and privacy for people with visual impairments. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*, 2016.

[6] Taslima Akter, Tousif Ahmed, Apu Kapadia, and Swami Manohar Swaminathan. Privacy considerations of the visually impaired with camera based assistive technologies: Misrepresentation, impropriety, and fairness. In *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility*, 2020.

[7] Taslima Akter, Bryan Dosono, Tousif Ahmed, Apu Kapadia, and Bryan Semaan. " i am uncomfortable sharing what i can't see": Privacy concerns of the visually impaired with camera based assistive applications. In *29th USENIX Security Symposium (USENIX Security 20)*, 2020.

[8] APWG APWG. Phishing activity trends report: 3rd quarter 2019. *Anti-Phishing Working Group. Retrieved April*, 2019.

[9] Shiri Azenkot, Kyle Rector, Richard Ladner, and Jacob Wobbrock. Passchords: secure multi-touch authentication for blind people. In *Proceedings of the 14th international ACM SIGACCESS conference on Computers and accessibility*, 2012.

[10] Natã M Barbosa, Jordan Hayes, Smirity Kaushik, and Yang Wang. "every website is a puzzle!": Facilitating access to common website features for people with visual impairments. *ACM Transactions on Accessible Computing (TACCESS)*, 2022.

[11] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*, 2014.

[12] Mark Blythe, Helen Petrie, and John A. Clark. F for fake: Four studies on how we fall for phish. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2011.

[13] Richard E Boyatzis. *Transforming qualitative information: Thematic analysis and code development*. sage, 1998.

[14] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*, 2018.

[15] Qian Cui, Guy-Vincent Jourdan, Gregor V. Bochmann, Russell Couturier, and Iosif-Viorel Onut. Tracking phishing attacks over time. In *Proc. of WWW*, 2017.

[16] Sauvik Das, Cori Faklaris, Jason I Hong, Laura A Dabbish, et al. The security & privacy acceptance framework (spaf). *Foundations and Trends® in Privacy and Security*, 5(1-2):1–143, 2022.

[17] Rachna Dhamija, J Doug Tygar, and Marti Hearst. Why phishing works. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, 2006.

[18] Bryan Dosono, Jordan Hayes, and Yang Wang. " i'm stuck!": A contextual inquiry of people with visual impairments in authentication. In *SOUPS*, pages 151–168, 2015.

[19] Vincent Drury and Ulrike Meyer. Certified phishing: taking a look at public key certificates of phishing websites. In *Fifteenth Symposium on Usable Privacy and Security (SOUPS 2019)*, pages 211–223, 2019.

[20] Adrienne Porter Felt, Robert W Reeder, Alex Ainslie, Helen Harris, Max Walker, Christopher Thompson, Mustafa Embre Acer, Elisabeth Morant, and Sunny Consolvo. Rethinking connection security indicators. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*, pages 1–14, 2016.

[21] Joseph L Fleiss, Bruce Levin, and Myunghee Cho Paik. *Statistical methods for rates and proportions*. john wiley & sons, 2013.

[22] Dan Geer. For good measure. *USENIX PATRONS*, page 72, 2020.

[23] Xiao Han, Nizar Kheir, and Davide Balzarotti. Phisheye: Live monitoring of sandboxed phishing kits. In *Proc. of CCS*, 2016.

[24] L. Hasty. Teaching tactile graphics, perkins school for the blind. https://www.perkinselearning.org/videos/webcast/teaching-tactile-graphics.

[25] Jordan Hayes, Smirity Kaushik, Charlotte Emily Price, and Yang Wang. Cooperative privacy and security: Learning from people with visual impairments and their allies. In *Fifteenth Symposium on Usable Privacy and Security (SOUPS 2019)*, 2019.

[26] Hilary Hutchinson, Wendy Mackay, Bo Westerlund, Benjamin B Bederson, Allison Druin, Catherine Plaisant, Michel Beaudouin-Lafon, Stéphane Conversy, Helen Evans, Heiko Hansen, et al. Technology probes: inspiring design for and with families. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, 2003.

[27] Ravi Kuber and Shiva Sharma. Toward tactile authentication for blind users. In *Proceedings of the 12th international ACM SIGACCESS conference on Computers and accessibility*, 2010.

[28] Victor Le Pochat, Tom Van Goethem, and Wouter Joosen. Funny accents: Exploring genuine interest in internationalized domain names. In *Proc. of PAM*, 2019.

[29] Eric Lin, Saul Greenberg, Eileah Trotter, David Ma, and John Aycock. Does domain highlighting help people identify phishing sites? In *Proc. of CHI*, 2011.

[30] Baojun Liu, Chaoyi Lu, Zhou Li, Ying Liu, Hai-Xin Duan, Shuang Hao, and Zaifeng Zhang. A reexamination of internationalized domain names: The good, the bad and the ugly. In *Proc. of DSN*, 2018.

[31] Alan Lundgard, Crystal Lee, and Arvind Satyanarayan. Sociotechnical considerations for accessible visualization design. In *2019 IEEE Visualization Conference (VIS)*, pages 16–20. IEEE, 2019.

[32] Daniela Napoli. Accessible and usable security: Exploring visually impaired users' online security and privacy strategies. 2018.

[33] Daniela Napoli, Khadija Baig, Sana Maqsood, and Sonia Chiasson. "i'm literally just hoping this will Work:'' obstacles blocking the online security and privacy of users with visual disabilities. In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*, 2021.

[34] Adam Oest, Yenganeh Safaei, Penghui Zhang, Brad Wardman, Kevin Tyers, Yan Shoshitaishvili, Adam Doupé, and Gail-Joon Ahn. Phishtime: Continuous longitudinal measurement of the effectiveness of anti-phishing blacklists. In *Proc. of USENIX Security*, 2020.

[35] Adam Oest, Penghui Zhang, Brad Wardman, Eric Nunes, Jakub Burgis, Ali Zand, Kurt Thomas, Adam Doupé, and Gail-Joon Ahn. Sunrise to sunset: Analyzing the end-to-end life cycle and effectiveness of phishing attacks at scale. In *29th {USENIX} Security Symposium ({USENIX} Security 20)*, 2020.

[36] Daniela Oliveira, Harold Rocha, Huizi Yang, Donovan Ellis, Sandeep Dommaraju, Melis Muradoglu, Devon Weir, Adam Soliman, Tian Lin, and Natalie Ebner. Dissecting spear phishing emails for older vs young adults: On the interplay of weapons of influence and life domains in predicting susceptibility to phishing. In *Proceedings of the 2017 chi conference on human factors in computing systems*, pages 6412–6424, 2017.

[37] Peng Peng, Chao Xu, Luke Quinn, Hang Hu, Bimal Viswanath, and Gang Wang. What happens after you leak your password: Understanding credential sharing on phishing sites. In *Proc. of Asia CCS*, 2019.

[38] Peng Peng, Limin Yang, Linhai Song, and Gang Wang. Opening the blackbox of virustotal: Analyzing online phishing scan engines. In *Proc. of IMC*, 2019.

[39] Yuji Sakurai, Takuya Watanabe, Tetsuya Okuda, Mitsuaki Akiyama, and Tatsuya Mori. Discovering httpsified phishing websites using the tls certificates footprints. In *2020 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, pages 522–531. IEEE, 2020.

[40] Pan Shi, Heng Xu, and Xiaolong Zhang. Informing security indicator design in web browsers. In *Proceedings of the 2011 iConference*, pages 569–575. 2011.

[41] Gunikhan Sonowal, KS Kuppusamy, and Ajit Kumar. Usability evaluation of active anti-phishing browser extensions for persons with visual impairments. In *2017 4th International Conference on Advanced Computing and Communication Systems (ICACCS)*, 2017.

[42] Abigale Stangl, Kristina Shiroma, Bo Xie, Kenneth R Fleischmann, and Danna Gurari. Visual content considered private by people who are blind. In *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility*, 2020.

[43] Christopher Thompson, Martin Shelton, Emily Stark, Maximilian Walker, Emily Schechter, and Adrienne Porter Felt. The web's identity crisis: Understanding the effectiveness of website identity indicators. In *Proc. of USENIX Security*, 2019.

[44] Ke Tian, Steve TK Jan, Hang Hu, Danfeng Yao, and Gang Wang. Needle in a haystack: Tracking down elite phishing domains in the wild. In *Proceedings of the Internet Measurement Conference 2018*, pages 429–442, 2018.

[45] Javier Vargas, Alejandro Correa Bahnsen, Sergio Villegas, and Daniel Ingevaldson. Knowing your enemies: leveraging data analysis to expose phishing patterns against a major us financial institution. In *Proc. of eCrime*, 2016.

[46] Yang Wang. Inclusive security and privacy. *IEEE Security & Privacy*, 16(4):82–87, 2018.

[47] Colin Whittaker, Brian Ryner, and Marria Nazif. Large-scale automatic classification of phishing pages. In *Proc. of NDSS*, 2010.

[48] Colin Whittaker, Brian Ryner, and Marria Nazif. Large-scale automatic classification of phishing pages. 2010.

[49] Guang Xiang, Jason Hong, Carolyn P Rose, and Lorrie Cranor. Cantina+ a feature-rich machine learning framework for detecting phishing web sites. *ACM Transactions on Information and System Security (TISSEC)*, 14(2):1–28, 2011.

[50] Yaman Yu, Saidivya Ashok, Smirity Kaushi, Yang Wang, and Gang Wang. Design and evaluation of inclusive email security indicators for people with visual impairments. In *2023 IEEE Symposium on Security and Privacy (SP)*, pages 1202–1219. IEEE Computer Society, 2022.

[51] Yue Zhang, Serge Egelman, Lorrie Cranor, and Jason Hong. Phinding Phish: Evaluating Anti-Phishing Tools. In *Proc. of NDSS*, 2007.

[52] Yue Zhang, Jason I Hong, and Lorrie F Cranor. Cantina: a content-based approach to detecting phishing web sites. In *Proc. of WWW*, 2007.

A Design Considerations

We designed a tool, GuardLens, to improve the accessibility of privacy/security cues of websites and help PVIs make more informed decisions while browsing websites online.

First, GuardLens provides easy access to privacy/security cues about a website (RQ1). This information is often inaccessible to PVIs but accessible to others almost instantly through a quick visual scan of a page (e.g., HTTPS lock icon, website search result ranking). The information otherwise readily provided by GuardLens is traditionally cumbersome to obtain or even inaccessible for PVIs, such as security certificate information [32, 33]. Motivated by prior work [10] and RQ1, one of our design goals was to *give users the ability to quickly obtain privacy/security information*.

Second, GuardLens hopes to help users with visual impairments protect against insecure websites (RQ2). Sighted users can rely on readily obtained privacy/security cues by simply glancing at a rendered page, enabling them to quickly take action to act on their privacy and security. For example, a quick glance may provide cues on whether a web page shows inappropriate images, what topic/genre the website or page is about (e.g., finance, news, shopping) and whether the page is out of context or is a click bait. In addition, with little additional effort, sighted users can also verify if links work or point to other website domains (e.g., via mouse-over), which can be helpful cues to detect phishing websites.

However, this is often not the case for PVIs. They use screen readers to navigate website content and often skip over large portions of text to prevent cognitive overload of information. However, doing so increases their likelihood of missing vital privacy/security related information [1]. Thus, obtaining privacy/security information about a website requires disproportionate effort on the part of PVIs. To this end and conforming with RQ2, our second design goal was to *provide equitable access to privacy/security-related information*, equipping users with useful information that could help protect them against privacy/security risks such as phishing websites. For instance, an attacker creates phishing (visually similar spoof) websites with a goal to trick users into entering personal information (e.g., account credentials, financial information). We assume that the attacker cannot alter information from trusted sources such as security certificates, domain registrations and Google search results.

Note that we conducted the formative study after developing GuardLens' initial version. We updated GuardLens tool design iteratively based on participant feedback from the formative study and the pilots.

B Formative Study

First, we conducted a formative study with five PVIs. Our **formative study** was motivated by prior work [10, 25, 41, 46] highlighting PVIs' needs for more accessible privacy/security

cues. The goal of this exploratory study was to understand the usage of GuardLens, as a technology probe, through 2-week field deployment. Each participant who completed the study for the full two weeks received a $70 gift card. We hypothesized that presenting a website's privacy/security cues in a non-visual format would help PVIs better assess the website. In particular, we explored two research questions: (1) what are the pros and cons in making website privacy/security-related information more salient to PVIs? (2) under what circumstances are privacy/security cues useful for PVIs?

To initiate the field study, we conducted a session with each participant to help them install the GuardLens browser extension. Due to the COVID-19-related social distancing guidelines, we conducted the study remotely via Zoom. After the initial session, participants used the system for two weeks as part of their regular browsing experience. Participants were asked to visit a minimum number of unique websites based on the screening survey. For example, if they claimed to visit 10-15 websites in the week prior to answering the screening survey, they were asked to visit at least 10 unique websites per week and half of the sites using GuardLens. After the 2-week period, we conducted 45 minute semi-structured exit interviews with participants. These interviews focused on the pros and cons of increased accessibility of privacy/security cues and whether the information provided by GuardLens was helpful. During the interview, we encouraged participants to share their experiences with GuardLens.

Participants found it difficult to access the security certificate of a website by clicking the padlock icon on the address bar. Therefore, GuardLens providing the security certificate information was useful. In addition, participants found three types of information from GuardLens most helpful: HTTPS encryption, external links pointing out of the website, and website owner. However, they also found the tool annoying because it would pop-up too frequently and it presented too much information. We used this feedback to improve the tool, for instance, by only showing the GuardLens pop-up when it detects important security issues (e.g., lack of HTTPS). We also added an option that allows users to choose specific privacy/security cues they want to see for a website. In addition, we made GuardLens more accessible, e.g., we improved the accessibility of the prompt dialog box (see Screen A in Figure 1) using an ARIA label.

C Pilot of Main Study

We pilot tested the main study with three PVIs, who self-identified as male, blind screen reader users. One of them did the earlier formative study. We followed the main study protocol and each pilot took about 1 hour. Each participant received a $30 gift card for completing the study.

Bird's eye view. Participants commended GuardLens' overview of privacy/security information of a website at one location. They said it saved them time compared to manually checking that information themselves. For instance, a participant said, *'When I am navigating without GuardLens, I don't have tool that tell me info about related links on the website and links to external websites. It gives me a quick bird-eye view of the website.'* Pilot participants found the following information from GuardLens most useful: website encryption (HTTPS), owner identity, and external links pointing out of the website. Participants assumed that if more links point out of the website, it may not be secure.

Feedback to improve GuardLens. Pilot participants reported that it was difficult to interpret the warning about 'owner identity unknown' because it only provided descriptive information but no actionable suggestions. We also observed that without GuardLens, participants applied strategies such as reading a website's URL character by character using their screen reader, and Googling unfamiliar websites to determine whether they are legitimate by checking their position in the Google search results.

Based on the findings from these pilots, we made several changes to GuardLens. We added two new information features to the system, namely, *domain age of website*, and *Google search results of a website*. The details about these cues were discussed in Section 3.2. We also added actionable suggestions for some of the cues, e.g., checking the website URL character by character as an actionable suggestion for the *'owner identity unknown'* cue.

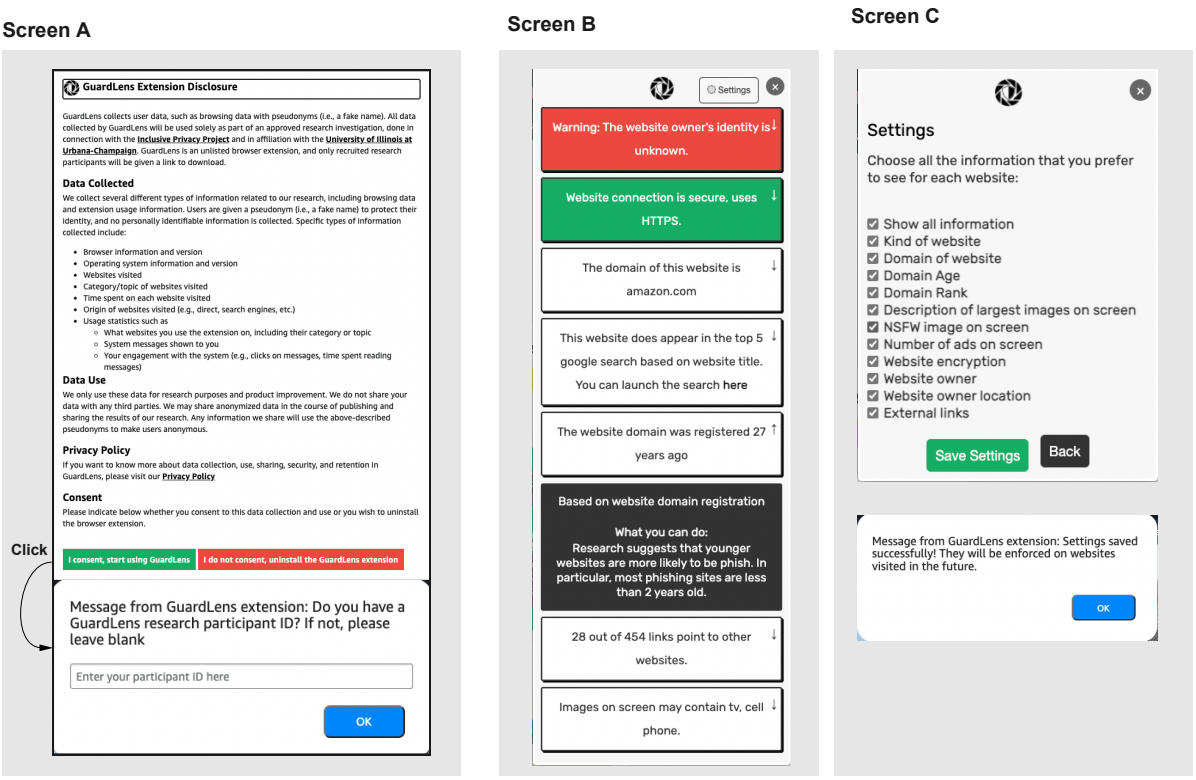


Figure 1: GuardLens UIs: (a) Screen A includes privacy disclosure and purpose of the study, (b) Screen B includes the main screen with privacy/security information blocks for a site being visited by a user and clicking into the arrow key of an information block will show tooltips and actionable suggestions, and (c) Screen C includes settings for the user to choose which information blocks to appear on GuardLens main screen as well as a confirmation page for saved settings. All the screens are marked up with the adapted information hierarchy and touch targets for screen reader accessibility.

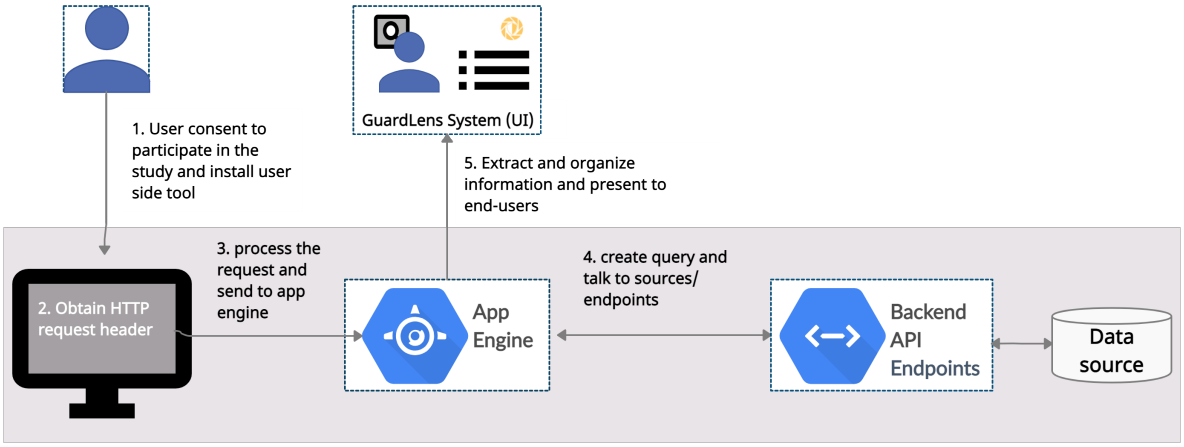


Figure 2: Workflow of GuardLens: upon user consent, GuardLens is triggered to send requests and build a channel between app engines (helpers) and external Backend API endpoints. App engines create queries, talk to data sources/endpoints, and present the structured information in the GuardLens UI for end users.

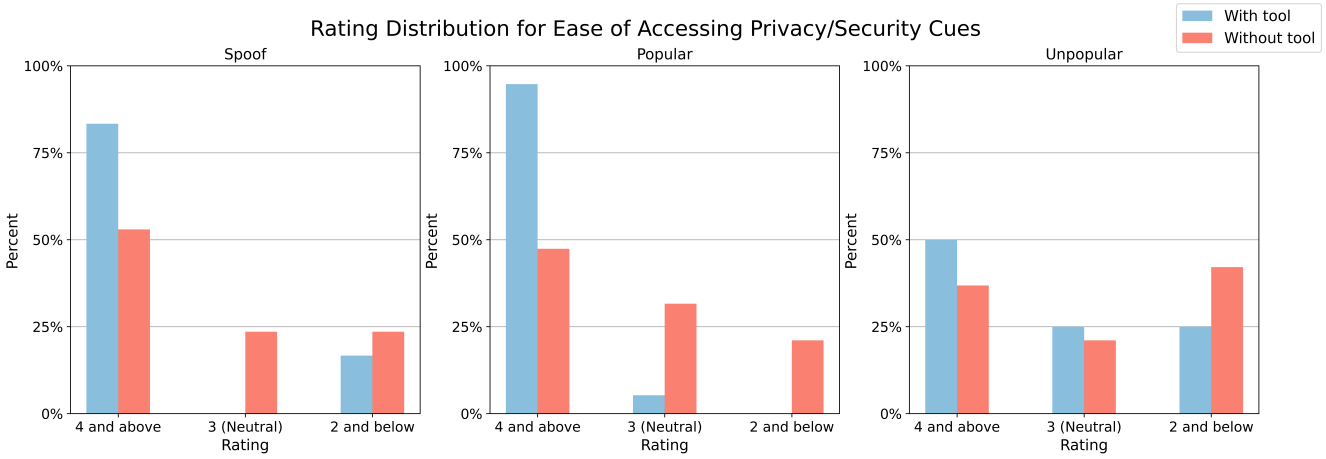


Figure 3: Participant ratings for ease of accessing privacy/security cues across three types of websites, spoof, popular, and unpopular websites, with and without GuardLens.

Table 1: The table shows most accessible privacy/security cues (in decreasing order) used by participants for three website types, i.e., spoof, popular, and unpopular, while browsing websites without and with GuardLens tool.

With/out Tool	**Spoof**	**Popular**	**Unpopular**
Without Tool	Read URL char by char Website footer Links HTTPS encryption in URL	HTTPS encryption in URL Website footer Links	HTTPS encryption in URL Website footer Links Website Accessibility
With Tool	Domain Age Domain Name	Domain Age HTTPS encryption (from tool)	HTTPS encryption (from tool) Domain Age Search Result Ranking

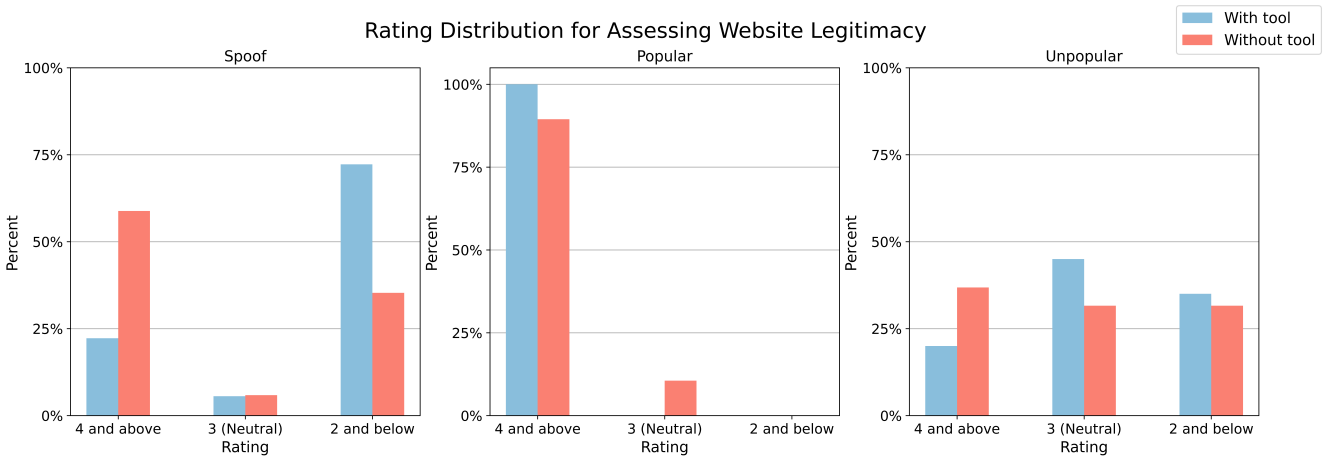


Figure 4: Participant ratings for website legitimacy across three types of websites (Spoof, Popular, and Unpopular) with and without the GuardLens tool.

Table 2: The table shows the most popular privacy/security strategies (in decreasing order) used by participants to assess the website legitimacy for three website types, i.e., spoof, popular, and unpopular, with or without GuardLens.

With/out Tool	**Spoof**	**Popular**	**Unpopular**
Without Tool	Familiarity with site Read URL character by character HTTPS encryption in URL	HTTPS encryption in URL Browsing content Familiarity with site	Google search by website title HTTPS encryption in URL
With Tool	Domain Age	Domain Age Google search result ranking (from tool) HTTPS encryption (from tool)	HTTPS encryption (from tool)

Table 3: Participant demographics (main study)

Participant ID	Order of Condition	Age Group	Gender	Self-Described Visual Ability	Assistive Technology Use	Education
P1	GuardLens First	45-54	Female	Blind	JAWS on laptop, VoiceOver on iPhone with Safari	Associate Degree
P2	Without tool First	55-64	Female	Blind Loss of Hearing	JAWS, VoiceOver, Refreshable Braille Display	Master's degree
P3	GuardLens First	25-34	Female	I can see lights, shadows, and objects very close to my face	JAWS, VoiceOver, ZoomText, Refreshable Braille Display	Master's degree
P4	Without tool First	25-34	Female	Blind	JAWS,Narrator, VoiceOver	Master's degree
P5	GuardLens First	35-44	Female	Blind	JAWS,NVDA, VoiceOver	Master's degree
P6	GuardLens First	18-24	Male	Blind	NVDA	Bachelor's degree
P7	GuardLens First	25-34	Male	Blind	NVDA,VoiceOver, Refreshable Braille Display	Bachelor's degree
P8	Without tool First	35-44	Female	Blind	JAWS,NVDA,VoiceOver, Refreshable Braille Display	Trade/technical /vocational training
P9	GuardLens First	18-24	Male	Blind	JAWS,NVDA,VoiceOver, Seeing AI, AIRA, Envision AI, ABB YY Fine Reader	Master's degree
P10	Without tool First	25-34	Male	Blind	JAWS,NVDA	Bachelor's degree
P11	GuardLens First	35-44	Male	I have retinal detachment	NVDA	No diploma
P12	Without tool First	35-44	Female	Blind	JAWS,NVDA,Narrator,VoiceOver, Refreshable Braille Display	Bachelor's degree
P13	GuardLens First	35-44	Female	Blind	JAWS	Master's degree
P14	Without tool First	25-34	Male	I'm diagnosed with RP (Retinitis Pigmentosa) with Maculer Degeneration and 100% blind	JAWS,NVDA,ORCA	Bachelor's degree
P15	GuardLens First	35-44	Male	Blind	JAWS	Master's degree
P16	Without tool First	18-24	Female	Totally blind except for light perception	JAWS, VoiceOver	High school graduate
P17	GuardLens First	25-34	Male	Retinopathy of prematurity, rop5; no light perception.	JAWS, VoiceOver, ABBYY	Professional degree
P18	Without tool First	65-74	Male	Blind	JAWS, Refreshable Braille Display	Professional degree
P19	GuardLens First	65-74	Male	Blind	JAWS, NVDA, Narrator, VoiceOver	Master's degree

Table 4: Websites from seven categories: finance, e-commerce, accessibility, news/media, education, healthcare, and productivity.

Website	Type	Genre
https://nfb.org/	Popular	Accessibility-related
https://aira.io	Popular	Accessibility-related
https://nytimes.com	Popular	News/Media
https://www.webmd.com	Popular	Health
https://www.target.com/	Popular	E-commerce
https://www.zapsend.co/index.php?//	Unpopular	Finance
https://yourcodercamp.com	Unpopular	Education
http://zolucky.com/	Unpopular	E-commerce
http://www.openoffice.org/	Unpopular	Productivity
http://audiobookbay.ws/	Unpopular	Audiobooks
https://www.amaZAUNN.com	Spoofed Amazon	E-commerce
https://www.enеffbee.org	Spoofed NFB	Accessibility-related

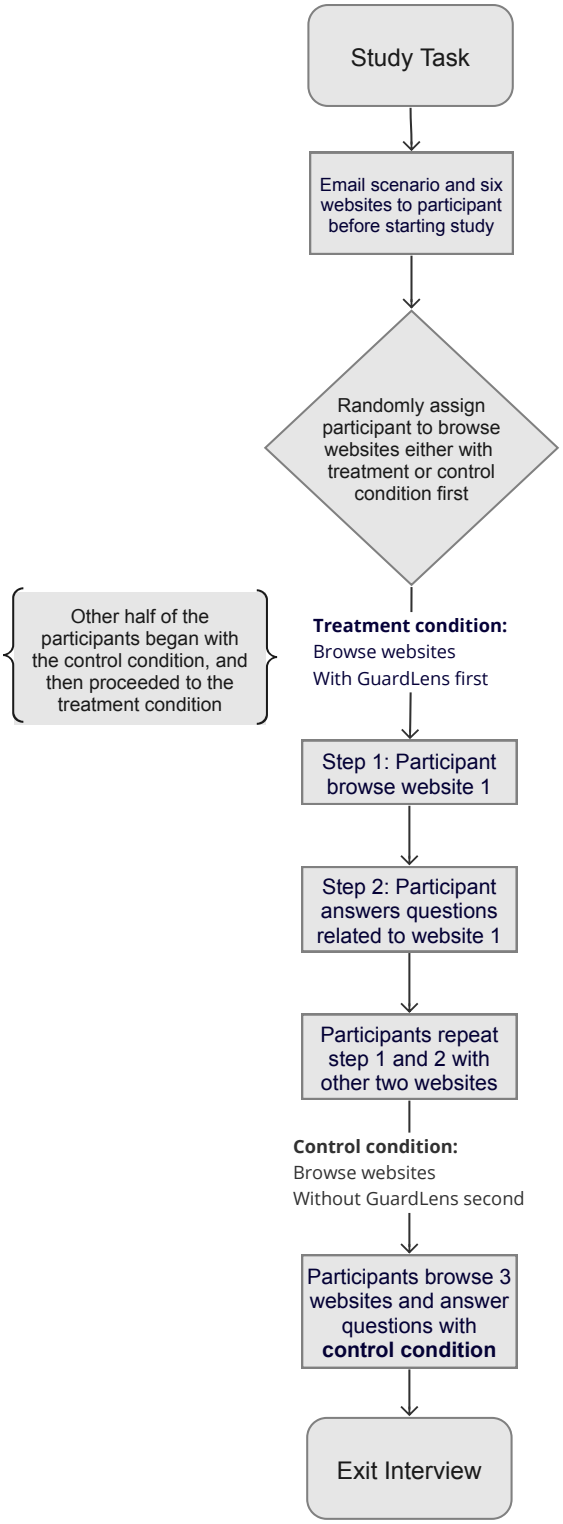


Figure 5: The main study design included the study task and the exit interview. In the study task, we emailed participants links to the websites along with a scenario. They visited various sites with and without GuardLens, unaware of site conditions, in a counterbalanced order.

Iterative Design of An Accessible Crypto Wallet for Blind Users

Zhixuan Zhou*
University of Illinois at Urbana-Champaign

Tanusree Sharma*
University of Illinois at Urbana-Champaign

Luke Emano
University of Illinois at Urbana-Champaign

Sauvik Das
Carnegie Mellon University

Yang Wang
University of Illinois at Urbana-Champaign

Abstract

Crypto wallets are a key touch-point for cryptocurrency use. People use crypto wallets to make transactions, manage crypto assets, and interact with decentralized apps (dApps). However, as is often the case with emergent technologies, little attention has been paid to understanding and improving accessibility barriers in crypto wallet software. We present a series of user studies that explored how both blind and sighted individuals use MetaMask, one of the most popular non-custodial crypto wallets. We uncovered inter-related accessibility, learnability, and security issues with MetaMask. We also report on an iterative redesign of MetaMask to make it more accessible for blind users. This process involved multiple evaluations with 44 novice crypto wallet users, including 20 sighted users, 23 blind users, and one user with low vision. Our study results show notable improvements for accessibility after two rounds of design iterations. Based on the results, we discuss design implications for creating more accessible and secure crypto wallets for blind users.

1 Introduction

Crypto wallets are an essential touch point for users to interact with blockchain and cryptocurrency technologies. These wallets are user interface wrappers over public/private key pairs that allow users to securely store, send, receive, and monitor their digital assets without mastering the underlying blockchain technology or running their own blockchain nodes [54]. In addition, they facilitate authenticating into and interacting with decentralized applications (dApps) [5, 16], and simplify the participation in the governance of Decentralized Autonomous Organizations (DAOs) [56]. In short, end-user use of cryptocurrency and blockchain technologies is synonymous with the use of crypto wallets.

Unsurprisingly, understanding and improving the end-user experience with crypto wallets has been the subject of much prior research [15, 25, 28, 59]. Yet, as is common with emergent and rapidly evolving technologies [20], ensuring the accessibility of crypto wallets has not been a focus of this prior effort. This lack of accessibility, in turn, effectively marginalizes blind users from participating in the emerging ecosystem of cryptocurrency and blockchain apps [18]. In response, Marta Piekarska, Director of Ecosystem at Hyperledger, highlighted the importance of accessible wallets for the blind community and how current wallets have not addressed their needs [27]. The National Federation of the Blind has, likewise, called for improving the accessibility of cryptocurrency technology [6], noting a lack of centralized oversight could cause poor accessibility outcomes.

Moreover, accessibility is closely tied to usability and security. Usability issues often disproportionately impact blind users [45]. For example, while all novices may be overwhelmed by the highly technical concepts foregrounded by crypto wallets (e.g., seed phrases, transaction gas fees) [3], novices who are blind face the additional burden of confronting these concepts with user interfaces that are inaccessible to screen readers. Likewise, security concerns are rampant in the Web3/crypto space, particularly phishing scams that trick users into sharing their private key or seed phrase with attackers [48] and sending their cryptocurrency assets to an attacker's address [37]. Blind users may be even more at risk than others: prior work has shown that accessibility tools can help blind users detect and protect themselves against phishing attacks [17], a tactic that cannot be executed when crypto wallets themselves are inaccessible.

We present a multi-phased, iterative re-design of a popular crypto wallet, MetaMask, to both examine and improve accessibility of crypto wallets. MetaMask is a non-custodial wallet, meaning that end users are responsible for managing their own private keys. We focused on MetaMask because it was the dominant wallet when we conducted this work in 2022, with 30 million monthly active users [52]. We designed the resulting wallet, iWallet, with "inclusiveness" in mind, aiming to improve accessibility for blind users. Specifically, our work was guided by the following research questions:

RQ1: What are the experiences of blind users with current crypto wallets?

*The first two authors contributed equally to this work.

RQ2: How can crypto wallets be made more accessible to blind users?

To answer RQ1, we collected and analyzed user reviews of 10 existing crypto wallets, performed a competitive analysis of the accessibility and other usability aspects of them, and conducted a usability study (N=18) of MetaMask (Version 10.23.2). Accessibility issues were common in all of the 10 popular wallets we analyzed. Eight of the ten wallets in our competitive analysis did not implement any accessibility features. The two that did (MetaMask, Coinbase) still had accessibility issues such as a confusing heading hierarchy, poor contrast between text and background colors, and a lack of keyboard navigation accessibility. In the usability study, we observed the behavior of 10 sighted users, 7 blind users, and one user with low vision while they performed basic tasks (e.g., creating an account, making transactions, importing an account) using MetaMask. We aimed to identify accessibility concerns and tasks that were disproportionately difficult for blind users to inform our later accessibility-centered redesign. We uncovered several accessibility issues, including unlabeled and poorly labeled buttons and a lack of confirmation notifications. These accessibility issues also exacerbated a number of correlated usability and security issues. For instance, when creating or importing a wallet account in MetaMask, users had to manually write down and type in their seed phrase, consisting of 12 automatically generated words — a challenge that was much more difficult for blind participants.

To answer RQ2, we followed an iterative design process [47] to implement and evaluate a more accessible version of MetaMask: iWallet. Based on our findings for RQ1, we designed iWallet with a focus on accessibility — touching on education, security, and usability. To evaluate our first redesign, we conducted a pilot study with a new set of 10 sighted and 8 blind participants. Building on their feedback, we iterated on our design and conducted a summative evaluation with a new set of 8 blind users. In our final design, we updated several features: we improved button labels, provided text summaries to improve the accessibility of video instructions, and prioritized downloading the seed phrase as a back-up option over manually writing it down. The summative evaluation confirmed that participants found iWallet more usable and accessible than the original MetaMask. On the System Usability Scale (SUS), participants rated iWallet much higher than MetaMask (81 vs 70, out of 100). Much of this improvement could be attributed to reducing complexity, improving ease of use, and reducing the need for prerequisite knowledge in their interactions with the wallet. Our blind participants, in particular, expressed positive feedback overall. Specifically, they appreciated: (i) the adequate labeling of buttons and web elements as well as the accessible secret recovery phrase[1] management process; (ii) being prompted to re-type and confirm the receiving address when sending cryptocurrencies to other accounts to ensure transaction security, since differentiating wallet addresses is harder for blind users; and, (iii) the accessible, video- or text-based explanations of technical crypto concepts (e.g., wallet address).

Our work makes two main contributions. First, through a competitive analysis of popular crypto wallets and a multi-phase user study, our work is the first to examine and improve the accessibility of crypto wallets for blind users, cataloging accessibility issues faced by blind users when using popular crypto wallets such as MetaMask. One concerning finding is that popular wallets such as MetaMask failed common accessibility standards such as WCAG. Second, through an iterative design process, we fix accessibility issues (e.g., by adding labels), introduce new accessibility designs such as downloadable, encrypted seed phrases, and provide key insights for researchers and practitioners on how to design more accessible crypto wallets for blind users.

2 Related Work

2.1 UX of Blockchain-Based Apps

User experience (UX) is a major challenge in blockchain-based applications, especially for non-technical users who find it daunting to understand the technical aspects of blockchain [30]. Additionally, the security of blockchain technology can be compromised if the UX is not designed with security in mind, leading to security breaches [34].

Decentralized Autonomous Organization (DAO), a disruptive advancement in blockchain-based applications, achieves algorithmic governance through smart contracts while heavily relying on human collaboration in decision-making [42, 44, 51]. DAOs face usability issues due to the complex nature of the underlying technology, particularly with respect to smart contracts. Users often encounter difficulties managing their tokens to participate in voting and proposals, which is considered challenging for less tech-savvy individuals who may not fully understand the process [33]. Non-fungible token (a.k.a. NFT), another popular blockchain application, also suffers from poor user experience, especially during the onboarding process, which can lead to loss of money and scams [53, 60]. Uniswap is a decentralized cryptocurrency exchange with two main features: cryptocurrency swapping and pooling cryptocurrency as liquidity. A research report indicated that the analytics features were not easily accessible to users navigating the Uniswap app, which could cause confusion and hinder user adoption [7]. Furthermore, users' mental models, influenced by traditional financial applications such as stock exchanges, could impact their perceived usability and user experience of this decentralized platform. In the case of blockchain-based gaming, players may find the technology not intuitive, with accessibility being a significant barrier [4].

2.2 UX of Financial Apps

Since crypto wallets could be considered a type of financial app, we also looked into the prior literature in financial apps,

[1] The secret recovery phrase in iWallet and MetaMask is equivalent to the seed phrase in other wallets. We use these two terms interchangeably.

identifying a number of issues related to user experience, including trust, security, usability, and design [14, 32, 38]. Akturan et al. conducted a user experience inspection of financial applications, focusing on mobile banking, online trading, and personal financial management [13]. Medhi et al. [43] examined mobile banking user interfaces in developing countries and proposed a framework that emphasized the importance of designing for users with limited literacy and numeracy skills to improve accessibility and financial inclusion. Teresa et al. [55] examined different methods and tools for evaluating the user experience of financial services and highlighted the challenges of conducting research in the highly regulated financial industry for many researchers. Several studies have investigated the relationship between trust and user experience in mobile banking using a combination of surveys and interviews. For example, it was found that trust was a key factor in user experience and that factors such as security, reliability, and transparency could have a significant impact on users' trust in mobile banking applications [62]. In addition, Wentz et al. and Goundar et al. [31, 61] have highlighted the lack of consideration for accessibility in designing various digital financial services.

2.3 Usability of Crypto Wallets

Cryptocurrency is becoming increasingly popular in recent years. According to an NBC News poll, one in five Americans has invested in, traded, or otherwise used cryptocurrency [24]. People hold cryptocurrency for investment purposes, to purchase goods including everyday items, and to learn more about crypto assets out of curiosity. Crypto wallets are software wrappers on top of private/public key pairs to facilitate easy interaction with the underlying blockchain [22]. However, usability issues prevent them from reaching the mass public [57, 63]. Researchers in the HCI community have tried to identify usability issues in crypto wallets [26]. For example, Moniruzzaman et al. adopted an analytical cognitive walk-through inspection with 5 participants, and found many crypto wallets lacked good usability in performing fundamental tasks [45]. From an end user's perspective, a blockchain usability report identified common usability issues of crypto wallets through a survey of over 200 crypto holders [23]. The report found that users had difficulty interacting with wallets and, in turn, the underlying blockchain. More than half of the users had at least one concern or problem with their transactions. Many users did not have full confidence in transactions, fearing that something might go wrong. The most reported issue was that users were not sure if a provided wallet address was accurate. Transaction fees were conceptually confusing to the users since the connection between fees and delivery times was often unclear. Similarly, 6,859 reviews regarding user experience of five mobile crypto wallets were identified and qualitatively analyzed by Voskobojnikov et al. [58]. Lack of guidance during the setup made it challenging to create a wallet. Qualitative quotes by interview participants were presented by Voskobojnikov et al. [57], pointing out usability issues of popular crypto wallets such as MetaMask: *"You have to enter a gas amount in some other currency that you have never heard called Gwei and then a lot of the times the recommended amount isn't enough."*

Numerous attacks have been conducted against the blockchain ecosystem [11], especially decentralized finance (DeFi), such as flash loans [50]. Such security vulnerabilities could often be attributed to usability issues. Mai et al. [40] found that users' misconceptions of cryptocurrency and the blockchain were associated with their inappropriate security and privacy practices. Compared to the large body of work devoted to understanding and addressing usability issues of crypto wallets, their accessibility is overlooked in prior literature, which we elaborate on in the following section.

2.4 Accessibility of Crypto Wallets

Accessibility in the context of crypto wallets has been discussed and promoted in earlier days. A blockchain workshop position statement highlighted that most crypto wallets relied heavily on visual elements, which made them difficult or even impossible for blind users to use [27]. Advocacy groups, such as the National Federation of the Blind, have responded to the accessibility challenges and called for increased attention to the accessibility of crypto technology. They argued that the decentralized nature of cryptocurrencies and the lack of centralized oversight may hinder accessibility for blind users [6]. However, to date, there has been a lack of academic work on the accessibility of crypto wallets. Thus, it is evident that there is a need for more focused research on the accessibility issues faced by blind users when using crypto wallets, as well as a need for the development of wallet designs that are accessible to all users, including those with disabilities. In this study, we aim to contribute to the limited literature in understanding and addressing accessibility issues of crypto wallets.

3 Empirical Analysis of MetaMask

No prior studies have examined wallet accessibility for blind users. To help fill the gap, we first analyzed 10 popular wallets in terms of their features and user reviews from three major platforms, i.e., Chrome Web Store (Chrome extension), App Store (iOS), and Google Play Store (Android). Details of this analysis are in Section A in the Appendix. Our competitive analysis allowed us to explore potential accessibility challenges which were then used to inform our study and redesign. In particular, we found common complaints about the lack of accessibility among these crypto wallets, such as poorly labeled buttons, and learnability and security issues.

To complement our aforementioned analysis, we conducted a user study with 10 sighted users, 7 blind users, and one user with low vision with MetaMask (Version 10.23.2). They were recruited from blockchain channels on Discord, Twitter, etc., as well as our participant pools of previous accessibility studies. Table 3 in Appendix shows the details of these participants

(M1-M18). We followed a similar procedure as in Section 5 for this exploratory user study, where we asked our participants to conduct a few tasks, such as creating a wallet account and sending some (testnet) tokens. Interviews were conducted before and after the tasks. The whole process generally took 1-2 hours. Blind users spent longer time on the tasks given the accessibility issues. Participants were given $30 as a compensation. Qualitative and quantitative data collected in this stage were used to inform our redesign to improve the accessibility and usability of MetaMask.

Our data came from participants' think-aloud responses and our observation notes during the tasks as well as the interview responses. The success rate of tasks and results from the SUS survey helped assess usability and accessibility for blind users. The educational aspect was measured through knowledge question (KQ) surveys and task success rates. We also examined tasks such as typing the correct receiving address and avoiding seed phrase disclosure to understand usable security implications.

3.1 Findings: MetaMask

The user evaluations of MetaMask revealed a number of issues about accessibility, security, as well as education about crypto literacy. On average, our sighted participants finished 8.6 of the 10 tasks. Blind users similarly finished 8.3 tasks, but the process was more cumbersome for them. It took sighted users 28.2 minutes on average to finish the tasks, while for blind users, the time increased to 47.9 minutes (about 70% longer than sighted users). A major reason for this time difference was the accessibility challenges encountered by blind users when using the wallet. Table 1 summarizes the quantitative results of our evaluations. We detail our findings next.

Accessibility. The majority of blind participants were more or less discouraged by accessibility issues, including unlabeled buttons and other web elements (e.g., input fields), lack of confirmations or notifications, incompatibility with screen readers, and the cumbersome process of dealing with the secret recovery phrase.

Buttons in MetaMask were not properly labeled to be readable by screen readers[2], according to many participants such as M13. This made wallet usage awkward for blind users. In the onboarding process, M11 failed to set up his password promptly since the password rule ("8 character min") was not readable by his screen reader. He also could not reveal the hidden secret recovery phrase in the onboarding process, since the button of "click here to reveal secret words" was hard to find and operate with a screen reader; thus he could not go to the next page without the help of our research team. During the transaction process, the field to enter the transaction amount was also not labeled (M11, M13), making the transaction a rather time-consuming process for the blind users. For some checkboxes, screen readers mistakenly announced them as unchecked even after the users checked them (M18).

Inconsistent notifications were raised as another accessibility issue in MetaMask. Sometimes, announcements were not provided after an operation was performed, e.g., copying the wallet address, or submitting a transaction. On the MetaMask wallet main page, the wallet address is provided. After one hovers over this clickable button with their mouse, a text popup would appear, saying "Copy to clipboard." However, if a blind user navigates to this button with their keyboard, the text will not be verbally announced. Many of our blind participants like M13 did not know how to copy the wallet address until being told by us to press the Enter key on the button. There was also no announcement after the wallet address was copied (M11), while for sighted users, there was a text popup "Copied!"

When verifying the secret recovery phrase during onboarding or importing accounts, blind users had to spend much time and energy confirming it word by word. In MetaMask, each secret recovery phrase is a random set of 12 words and users need to select the words in the correct sequence to verify it. As in M18's case, when confirming the secret recovery phrase, she needed to check the upcoming word in the original phrase, and go through the shuffled words to find the matching one — this process was cumbersome using a screen reader. M15 also complained that this confirmation process took a lot of energy. Some blind users such as M11 and M13 did not want to go back and forth to confirm the secret recovery phrase and chose to use the "remind me later" option to skip the process, which could become a significant security risk as MetaMask did not provide an intuitive way for them to go through this process again later in use. M13 skipped the process and could not import her wallet later since she did not have the secret recovery phrase. M11 downloaded the secret recovery phrase in plain text in a file, which could be easily left in the wrong hands. He felt MetaMask should provide more secure and accessible options for storing the secret recovery phrase.

Many expressed that accessibility issues could also lead to security problems. For example, M11 explained how improper button labeling could lead to security and trust issues for blind users, *"With some of the functionalities unclear to screen reader users, it raises trust concerns in that I'm afraid to set off some unknown function that could negatively impact my account. For example, I might accidentally click a wrong button which is not labeled, and reveal my account. It could be a real problem if I'm on public channels."* This chain effect of accessibility issues and subsequent fear of accidentally revealing important private information can significantly limit their ability to engage with the wallet and the crypto space.

While the aforementioned issues were spotted in an earlier version of MetaMask (Version 10.23.2), the more recent version (Version 10.25.0) made some color changes in its light mode to improve color accessibility. However, other accessibility issues we identified still persisted. The design changes

[2] Our participants mostly use JAWS, NVDA, and Voiceover (only available on Mac systems).

we made later were still lacking in the latest version of MetaMask. We were not directly collaborating with MetaMask, but plan to share our findings and redesigns with MetaMask.

Education & Learnability. Except for M3 who was familiar with the concepts of crypto wallets before the study, all other participants found MetaMask confusing due to its many complicated concepts, such as secret recovery phrase, private key, gas fee, and main vs. test networks. For instance, many reported a lack of education for gas fees. Some participants, like M6, suspected that the gas fee might be a transaction fee equivalent, but none of them were sure about this concept. M1 was confused about what percentage of the transaction amount she should pay as the gas fee. Similarly, many participants did not understand the meaning and importance of the secret recovery phrase even after the study. These concepts are common in crypto wallets, which could give novice users an extra barrier when using them. M16 thought the option of skipping secret recovery phrase confirmation ("Remind me later") in the onboarding process diluted the education on this concept, and he was no longer sure if it was important ("Skippable things are not important").

MetaMask often failed to provide explanations on crypto concepts, which could be challenging for blind users. For example, the shortened wallet address (e.g., 0x056...8089) was there without further explanation, with many blind participants not knowing it was the wallet address. Sighted users could infer the role of the string, i.e., they inferred the wallet address string was the address after seeing it: *"I guess it's the wallet address. Addresses usually look like this."* However, blind users found more difficulty doing so since they could not visually see the wallet address to infer what it was.

Moreover, our participants tended to skip educational videos out of their user habit with apps. Blind users like M16 preferred text-based instructions, since they were easier to read, more accessible, and more time-saving than videos. The education provided by MetaMask was associated with a small improvement in the number of KQs answered correctly. The participants answered only 0.3 more questions correctly in the post-study KQ survey than in the pre-study one.

Usable Security. A few participants (both sighted and blind users) mentioned their concerns about sending crypto assets to wrong receiving addresses, since in MetaMask, there was not a confirmation page asking users to double-check their transaction details such as receiving address and amount. M9, who manually typed the receiving address during the transaction task, expressed the fear of typing a wrong address. Several blind participants typed their own wallet address instead of the one provided by the research team, and corrected it after being reminded by the research team.

Disproportionately Impacted Blind Users. Our sighted and blind participants experienced security and learnability issues in MetaMask, such as uncertainty about transaction accuracy and a lack of explanation for crypto concepts. These issues were in part because they were novice users of cryptocurrencies and crypto wallets. However, these issues could affect blind users even more. For instance, without accessible (visual) cues, it was harder for blind users to tell different addresses apart. Moreover, inaccessible features such as unlabeled buttons made it a rather cumbersome process to use MetaMask for blind users.

4 Our Redesign of MetaMask

Previous studies in crypto wallets have primarily focused on investigating the usability challenges [45] and security perceptions [40] of blockchain technologies. However, there remains a lack of clarity regarding how to design crypto wallets to accommodate a broader user population, which includes novice users with limited cryptocurrency literacy, and individuals with visual impairments. Our competitive analysis of multiple popular wallets (described in Section A in the Appendix) and user evaluations of MetaMask (detailed in Section 3) have revealed several limitations and shortcomings that informed the areas of improvement for accessibility as well as education and usable security. While blind users were disproportionately impacted by the design flaws, we utilized the results to guide our accessibility-centered designs. To explore ways to address these issues, we employed an iterative design process to implement and evaluate our redesign ideas.

4.1 Crypto Wallet Redesign Considerations

Existing crypto wallets are often cumbersome to use for blind users due to a number of reasons, such as inadequate labeling, core functions which require great cognitive effort from blind users (e.g., secret recovery phrase management), and ineffective and inaccessible learning resources. Thus accessibility became a pivotal design consideration in our study. To enhance accessibility, we aimed to label buttons and other web elements adequately, organize them into a clear hierarchy, and use a combination of colors with sufficient contrast to make the content more distinguishable for users with low vision. In addition, we considered streamlining the cumbersome secret recovery phrase management during the processes of onboarding and account importing to improve usability while potentially enhance accessibility for blind users.

Design consideration 1: We improved the accessibility of crypto wallets by labeling buttons adequately, organizing web elements, and simplifying complicated tasks such as secret recovery phrase management.

While crypto concepts were found harder to grasp for blind users without visual cues, we aspired to present design features that improved both accessibility and learnability of crypto wallets. To enhance the accessibility of educational resources on crucial concepts and terminologies in cryptocurrency, such as gas fees and secret recovery phrase, we aimed to incorporate intuitive onboarding and transaction processes featuring a well-informed navigation with just-in-time video and text instructions. The embedded instructions and guidance would potentially eliminate the need for users to switch

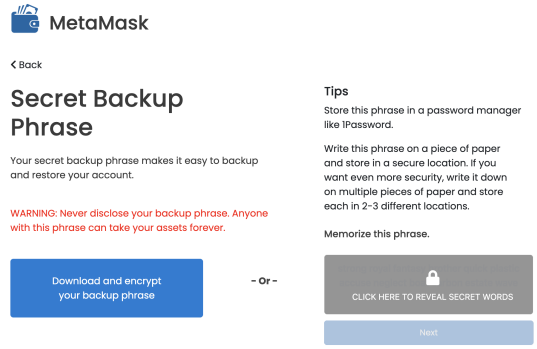


Figure 1: Downloadable encrypted secret recovery phrase for seamless management.

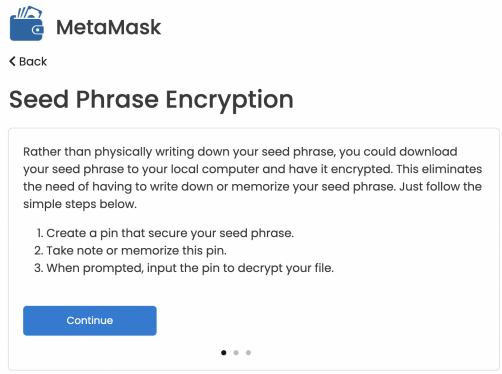


Figure 2: A PIN code to encrypt the secret recovery phrase before downloading it to enhance security.

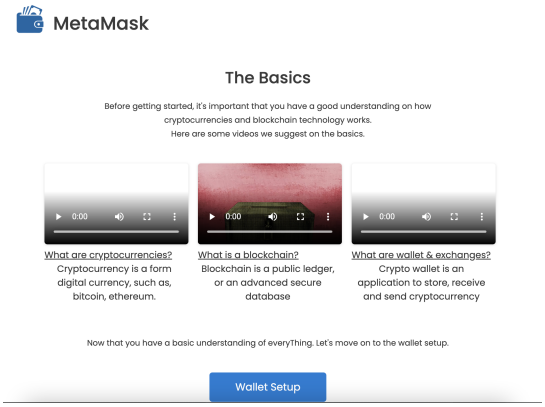


Figure 3: A dedicated education page with accessible videos and summative text to educate users on blockchain and wallet basics during onboarding.

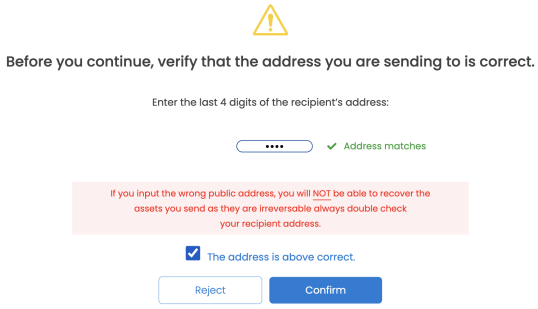


Figure 4: A confirmation page for users to double-check the receiving address by re-typing the last 4 digits to improve transaction security.

to a separate help center or additional web pages, which could pose a challenge for blind users. The instructional videos and text would educate novice users about crypto concepts such as secret recovery phrase and how to securely use the wallet.

Design consideration 2: We embedded just-in-time educational resources including videos and text to help blind users understand critical concepts and explore wallets.

Finally, we considered incorporating the feature that allowed users to verify the last four digits of the recipient address in the transaction process. This was motivated by the results of our user study with MetaMask, where participants strongly suggested adding a confirmation page when financial assets were at stake. The blind participants often could not tell different wallet addresses apart, which made them prone to the risk of sending crypto assets to the wrong addresses. Such security features could enhance the overall security of users' assets, especially for blind users.

Design consideration 3: We added security features to help reduce risks for novice users especially for blind users.

4.2 Redesign: iWallet (V1)

We used MetaMask, the most popular crypto wallet, as the template for our redesign in terms of accessibility, education, and usable security for blind users.

Accessibility. To improve the accessibility of MetaMask, we took inspiration from Universal Design (UD), a framework for inclusive design, providing principles for accessible experiences in both physical and digital domains [39]. These principles include equal conditions, flexibility, simplicity, tolerance for error, and reduced physical effort. UD implementation is guided by best practices and accessibility guidelines such as Web Content Accessibility Guidelines (WCAG) [8, 36] and

Americans with Disabilities Act (ADA) [1]. More specifically, we first improved labeling for screen readers and darkened texts for better contrast toward meeting the WCAG standards [8], which MetaMask failed to meet. The ARIA [2] specification, which provides a framework to improve the accessibility and interoperability of web content and applications, was also implemented for additional contexts for screen readers. To help blind users be aware of their status in operations, we provided notifications in pop-up windows after they finished each task, e.g., submitting a transaction. We further designed the function of downloading and uploading an encrypted version of the secret recovery phrase for better accessibility and security for blind users (Figures 1 and 2), as the process of writing and typing in the phrase was cumbersome and error-prone for these users.

Accessible Educational Resources. Our redesign in the education aspect utilized two types of media: video and text. Figure 3 displays an example of the redesign aimed at educating users on wallet usage and crypto concepts. Specifically, we added a dedicated page at the beginning of the onboarding process with three informational videos to give users a general understanding of blockchain, cryptocurrency, and crypto wallets/exchanges. Note that the summative text under the videos was implemented in the second design iteration. We also provided a video during the transaction process to explain the concept of gas fee, which was found confusing in previous research [23] and our evaluation of MetaMask. Some of our blind participants in the MetaMask evaluation expressed that they skipped videos because they were too fast to grasp, and they would prefer text. In response, we added text explanations using metaphors throughout the wallet pages. For instance, we used bank account number to explain a wallet address and bank password to explain a private key. To educate users about gas fee and its impact on transaction time in an intuitive, direct-manipulation manner, we designed a gas fee slider showing the positive relationship between gas fee and transaction speed, and allowing users to choose between low, average, or high fees for estimated transaction times of 45, 30, or 15 seconds, respectively.

Usable Security. To improve transaction security, especially for blind users, we designed a dedicated address confirmation page (Figure 4), asking users to re-type the last four characters of the receiving address; only if there is a match, users can proceed with the transaction. This feature was designed to help ensure crypto assets were sent to the right address.

4.3 Pilot Study Results of iWallet (V1)

We conducted a formative pilot study with 10 sighted users and 8 blind users (W1-W18) to get feedback about our initial redesign. We followed the same study process as in the evaluation of MetaMask. Main pilot results are summarized below but detailed in Section C in the Appendix.

The receiving address confirmation was regarded as a useful and accessible security feature. Our embedded education was deemed useful by participants, though the blind users expressed a preference of text, which was more accessible for them, over video instructions. While accessibility was greatly improved in iWallet, leading to lower task completion time than MetaMask, several accessibility challenges remain: (1) Blind users wanted explicit text explanations beside crypto concepts such as wallet address since they could not visually infer their meaning as sighted users; (2) Blind users expected more information (e.g., what is the next page for) in button labeling to assist their navigation; (3) Blind users wanted text summaries to supplement videos so that they could skip the often inaccessible videos without missing important information; (4) We failed to prioritize the option of uploading the secret recovery phrase in the account importing process, leading to task failure. Toward addressing the original accessibility issues and the ones in our initial redesign, we focused on accessibility in the second design iteration, which we elaborate next.

4.4 Redesign: iWallet (V2)

Based on the evaluation of iWallet (V1), our primary design goal was to enhance its accessibility. We provided more thoughtful button labeling in a hierarchical manner to mitigate physical and cognitive efforts, and brief descriptions of buttons and web elements to help blind users understand what information was contained in them. We further improved the consistency of page layout and prioritized the download/upload option of managing the secret recovery phrase during account creation and importing, aiming to make this accessibility design better received.

We also made education more accessible for blind users. To cater to user preferences and support their accessibility needs, we added text summaries to supplement educational videos, as some participants indicated that they preferred text instructions, which were more accessible with a screen reader, over videos. For example, under the video on cryptocurrency, we provided a text summary: “Cryptocurrency is a form of digital currency, such as Bitcoin and Ethereum (ETH).” Additionally, we made a number of changes to mitigate confusing points of the user interface that our blind users expressed during the pilot study. For example, MetaMask allows for the management of multiple “accounts,” i.e., public/private key pairs. By default, MetaMask refers to the wallet address of a user’s default account as “Account 1.” However, the concept of a wallet address as an “account” was confusing to users. While sighted users were able to infer the wallet address string to be an address, blind users found difficulty figuring it out. Thus we changed the heading from “Account 1” to “Wallet Address” to be more in line with the educational content we showed users on wallet setup. Similarly, we provided a text explanation for ETH, indicating it was a cryptocurrency.

4.5 Wallet System Design and Implementation

We developed our crypto wallet using React JS, and tested it on Chrome and Firefox to ensure its functionality. The backend of the wallet is capable of handling various functions, such as updating users' transaction history when they buy or sell cryptocurrency, facilitating local storage for data requests from APIs, and including general helper algorithms to streamline the redesigned features (e.g., gas slider design, downloadable encrypted seed phrase). Our wallet workflow includes essential features such as account creation, secret recovery phrase encryption, and gas fee adjustment. To save users' transaction activity, our crypto wallet app engine interacts with storage hosted on MongoDB. Whenever users' activity or queries are received, the wallet database updates the state to reflect the current activity on their end. Note that our crypto wallet is an unlisted browser extension, and only selected participants were asked to perform tasks during the study. Our wallet code is available on GitHub[3].

5 User Study: iWallet (V2)

In the main user study, we aimed to evaluate our proposed accessibility features after two design iterations. Below, we provide a detailed description of the recruitment process, experiment setup, and data analysis methods.

5.1 Recruitment

To evaluate iWallet (V2), we conducted a user experiment with novice blind crypto users (N=8, W19-W26). We defined novice users as those who may have heard of cryptocurrency but have not traded or used them yet, or those who only had experience with centralized exchanges (CEXes) such as Coinbase and Binance, but had no or little experience with non-custodial wallets such as MetaMask. The screening survey included questions asking about potential participants' experience with cryptocurrency, exchanges, and wallets, as well as demographic information such as gender, age, educational level, and country. Given the focus on accessibility of our wallet (V2), we specifically sought out to recruit blind participants. To this end, we asked about participants' visual acuity and screen reader(s) they used.

The participants were recruited from various cryptocurrency channels and forums, including Discord, Twitter, and Reddit, as well as previous participant pools of our accessibility studies. Ultimately, our goal was to recruit a diverse group of participants with varied demographic characteristics, as presented in Table 3 in the Appendix. All explicitly indicated an interest in cryptocurrencies and in using crypto wallets.

The participation in our study was completely voluntary, and participants were allowed to withdraw at any time. The participants received $30, in form of Amazon gift cards, as a compensation. The whole study lasted 1-2 hours for our participants, and blind users tended to spend more time finishing the study. The study was IRB approved.

[3]Wallet code: https://github.com/AccountProject/Wallet_App

5.2 Experiment Setup

Procedure. We started by conducting brief semi-structured interviews with the participants to gather insights about their prior experience with cryptocurrency and crypto exchanges, if any. After answering six knowledge questions (KQs), which were used to assess their crypto knowledge, they were assigned several tasks regarding crypto assets management and transaction. After finishing the tasks, we asked the participants KQs again to see if our education was effective and well received. We then asked participants to fill in a SUS questionnaire to evaluate the usability of the wallet. Finally, exit interviews were conducted to obtain participants' overall experience of using the wallet, including perceived accessibility and general usability. Suggestions for future wallet design especially regarding accessibility were also collected. We conducted all user experiments via Zoom, and recorded the interviews and tasks, upon consent, for further analysis. Below we detail the study design.

Exploratory Interview. After introducing our study to the participants, we asked about their experience and knowledge of cryptocurrency and its underlying infrastructure, i.e., blockchain. Most of our participants were truly novice crypto users, who had only heard of cryptocurrency, and we kept the interviews with them short. For those who had used crypto exchanges before, we asked about their general experiences with exchanges, especially regarding the accessibility aspects. We also asked them what general and accessibility features they would expect if they needed to use a crypto wallet to trade cryptocurrency.

Tasks. Before tasks began, knowledge questions were asked in a survey to assess participants' initial crypto knowledge, as a reference to that after performing tasks. The six knowledge questions (KQs) were multiple choice questions about core wallet concepts, i.e., token names (ETH), wallet addresses, seed phrases, account security, transactions, and gas fees.

Each participant was asked to perform several tasks with our wallet. The wallet has been pre-installed on the research team's local computer, and participants were given access to control this computer remotely to perform the tasks, which was a function afforded by Zoom. They were asked to think aloud and answer questions during the study. We recorded the participants' process of performing the tasks for later analysis.

The tasks (N=4) and sub-tasks (n=9) required to finish each task were revised from [45], and are listed below.We specifically sought out to observe how accessible our wallet was to novice/blind users, and identify any accessibility challenges.

- **Task T1**: Configuration - Creating a new account within the wallet (sub-task t1). Participants can optionally watch educational videos to understand the concepts of crypto, blockchain, and wallet.

- **Task T2**: Checking wallet address, (receiving test ETH), and checking wallet balance. We ask participants to provide their wallet address to us (sub-task t2), and check test ETH balance after receiving it (sub-task t3).

- **Task T3**: Spend/Transfer - Making a transaction of 1 test ETH to the research team. Thus, this task involves finding the transaction functionality (i.e., Send) (sub-task t4), entering information such as receiver's address provided by the research team (sub-task t5) and ETH amount (sub-task t6), submitting the transaction (sub-task t7), and expressing when they think the transaction is confirmed (sub-task t8).

- **Task T4**: New device scenario - Imagining using a different device and importing the existing account. We ask participants to refresh the wallet page (get back to the onboarding phase), and ask them to import their existing wallet into the "new" device (sub-task t9).

After using the wallet, we asked the participants to answer the KQs again to evaluate their crypto knowledge. The SUS survey was filled out to evaluate the usability of the wallet. In addition to the 10 items implemented in the SUS survey, we additionally contained 4 wallet-specific usability questions concerning the onboarding process, wallet address, transaction process, and gas gee.

Exit Interviews. In the exit interview, we asked participants for their opinions on various aspects of the wallet, including overall experience, accessibility, learnability, security, privacy, etc. We also asked for design suggestions from them to further improve our wallet, especially in terms of accessibility.

5.3 Data Analysis

We performed both qualitative and quantitative analysis on our collected data from interviews, task observations, and surveys.

Qualitative Analysis. Two authors independently coded the transcripts of the conversations during the experiments and interviews as well as observational notes, and met regularly to discuss. Through thematic coding [29], themes started to merge and brought us back to the transcripts to find more data for them. We used XMind [10], a mind mapping tool, to arrange and organize codes and corresponding quotes into a hierarchy of themes. After several iterations of analysis, we arrived at the current findings. We use observational notes and participant quotes to illustrate our points. All quotes have been anonymized to protect privacy of the participants.

Quantitative Analysis. Recorded videos of the participants performing assigned tasks were analyzed to measure task success rates and task completion times. Surveys, including two KQ surveys and one SUS questionnaire, were statistically analyzed to understand educational effect and user experience of the wallet.

6 Findings: iWallet (V2)

Our accessibility features were greatly improved compared to MetaMask and the previous version of iWallet (V1). iWallet (V2) outperformed MetaMask in terms of SUS score, task success rate, task completion time for blind users, and improvement in the number of KQs answered correctly after usage (see Table 1). The security and education features also helped blind users interact with crypto wallets smoothly and safely. Our participants provided rich qualitative insights regarding our accessibility and security improvement.

6.1 Accessibility

Our improvement in accessibility resulted in a higher SUS score (81), compared to 70 for MetaMask and 65 for our previous iteration. Moreover, none of our participants encountered any difficulties in completing the sub-tasks. Our blind participants spent 37.8 minutes on the tasks, compared to 40 minutes in V1 and 47.9 minutes in MetaMask.

Enhancing Accessibility for Equitable Use. Accessibility was key to blind users' interaction with crypto wallets and financial apps in general, but was frequently overlooked. Our participant, W26, reported her experience of trying multiple centralized and decentralized crypto exchanges before settling down to Robinhood, which worked well with iPhone and Voiceover. She added that Coinbase, another crypto exchange, was not accessible, leading to a lack of confidence in using the platform: *"It's confusing, not very friendly. I cannot get through it with confidence. If I accidentally hit a button When dealing with finances, I don't know if I will get into an area that I cannot get out of. Buttons should be labeled correctly instead of just reading 'button' or 'link'."* After using our wallet, she thought it well met her accessibility needs, making it possible and enjoyable for her to interact with crypto wallets. Many of them did not expect crypto wallets to be accessible, and felt it a pleasurable process to use our wallet. W20 expressed appreciation for our accessible wallet and inquired about its availability in the Chrome Web Store and Apple App Store for daily use.

Adequately Labeled Buttons for Intuitiveness. Our participants provided positive feedback regarding the clear and informative button labeling (W20, W21, W22, W23, W25, W26). Participant W23 highlighted the usefulness of the descriptive button labeling in guiding the completion of tasks, stating, *"I found the button names to be descriptive compared to other apps I usually use. When I complete a task and press the next button, I'm quite sure what I'm going to see next, like, the next is X or Y."* W25 similarly praised the well-labeled buttons: *"The buttons and input fields are pretty well labeled."*

Onboarding/Portability with Low Physical Effort. Our participants, e.g., W21, indicated experiencing no accessibility issues when making transactions and found the user interface workflow to be straightforward. W19, W20, W21, W23, and W26 found the streamlined onboarding process – whether creating a new wallet or importing an existing one –

Wallet	SUS score (out of 100)	Task success rate (%)	Task completion time (min)	Improvement in KQs
MetaMask	70	86	28.2 (sighted), 47.9 (blind)	+0.3
iWallet (V1)	65	79	26.1 (sighted), 40.0 (blind)	+0.7
iWallet (V2)	**81**	**100**	**37.8 (blind)**	**+1.3**

Table 1: Quantitative results of our evaluations.

to be effortless. W19 expressed that *"the onboarding process is pretty easy to follow,"* while W23 echoed her opinion, *"Account creation is intuitive and the steps are straightforward."* W26 was particularly intrigued by the onboarding process, which introduced her to the new concept of secret recovery phrase – something she had never encountered in her previous experience with Robinhood. Most participants preferred the option to download the encrypted secret recovery phrase over the option to write it down manually (W20, W22, W23, W24, W25, W26). W22 opted to upload the secret recovery phrase to verify the account, explaining that *"selecting the words takes too much time, and uploading seems less time-consuming."* W20, who chose to write down the phrase, had difficulties importing his wallet due to not inserting space between words while typing them manually. He stated that the process would have been easier if he had chosen the option to download the phrase. After the researchers consistently prioritized the download/upload option in the onboarding/importing process, none of the participants in this round had confusion as observed in the pilot study.

Enhancing Accessibility Enhances Security. The downloading option of managing the secret recovery phrase was also regarded as more secure since it was encrypted by a pin (W24). W23 further expressed his security concern of the writing option: *"If we are outside, there would be privacy issues. When we write it down in a mobile phone or computer, other people can hear it. Not all screen reader users use headphones."* Interestingly, several participants downloaded and wrote down the secret recovery phrase at the same time to avoid losing it (W21, W22, W24). This could be due to blind users' typically cautious behavior when using a new app.

The only remaining accessibility improvement recommendation pertained to ensuring seamless auto-focus on popup windows for screen reader users (W22, W24, W25, W26). According to W25, who used JAWS as the screen reader, the confirmation notification after the transaction was helpful for blind users; however, the appearance of popups was not spotted by the screen reader timely. As a result, she had to navigate to the bottom of the main page before finding the popup window. W24 suggested using dialog boxes rather than pop-ups to facilitate easier navigation for blind users.

6.2 Accessible Crypto Wallet Education

It was commonly acknowledged by our participants that having sufficient knowledge of crypto wallet was crucial for blind users since they found difficulty relying on visual cues to infer meanings of new concepts and inform their security decisions, echoing with our previous user evaluations. Before the tasks, most of our participants did not know basic concepts in crypto wallets such as secret recovery phrase and wallet address, as shown in the first KQ survey. W24 explicitly indicated that she was *"not sure what secret recovery phrase and wallet address meant."* W19 was on-boarded to a crypto wallet by her brother, but did not use it ever, as she found it conceptually harder to use than traditional financial apps. She added, *"Blockchain concepts and terminologies are a totally different world."* Before the experiments, our participants answered 3.3 questions correctly out of a total of 6 on average. Following the experiments, the number increased to 4.6. After using our wallet, 10 more participants (from 6 to 16) correctly answered the question on wallet address, compared to an increase of two for the baseline MetaMask. Such evidence demonstrates the efficacy of our accessible education designs, including rich instructional videos and concise text summaries as suggested by blind users in the pilot study, as well as text explanations of wallet address, ETH, etc.

Crypto Knowledge. The instructional videos were effective in familiarizing some users with crypto concepts (W21, W22, W23, W25, W26). W22 found the videos useful for learning how to create and use a new wallet: *"The information was totally new for me. That's why I watched the videos."* W25 intended to watch all the videos to *"make sure what it is, in case doing something incorrectly."* She thought the videos were concise and taught her interesting concepts such as mining. W26 thought the videos were of good length and helped explain things, such as what secret recovery phrase really meant. W24 found the videos easy to understand, played at reasonable speed, jargon-free, and informative. On the contrary, W20 did not watch any videos in the onboarding process and had difficulty understanding crypto terminologies such as the secret recovery phrase in subsequent steps. He acknowledged that he would have watched the videos if he knew there were related operations afterward. Additionally, some participants, such as W24, preferred to read the text summaries of the video content rather than watch the videos themselves.

By explicitly explaining the "Wallet Address" right above the string and putting the text "ETH is a digital currency" below the ETH balance, we enabled our participants to locate them promptly (W19, W20, W24). In contrast, there were no such text cues in MetaMask and our previous version, making

it harder for blind users to infer their meanings. Many of our participants, including W19, indicated using the explicit text evidence to find the wallet address.

Goal-Directed & Engaging Learning. The gas fee slider was deemed as an intuitive way to adjust gas fee by our participants (W19, W20, W23). W20 thought it was a nice design to have in the wallet. He further explained that he would select the low option when sending money to friends and the high option if he was sending money for business or critical transactions to make it more timely. W25 went with the default gas fee, since she *"didn't know enough about it [gas fee] to make it go through."* W22 also chose the default gas fee because she thought *"the system default would be a good option, and users tend to think the default one is the best option for them."* W24 further suggested a design to show how many people were choosing each option to inform new users' decisions and increase their trust of the slider as a social navigation feature.

6.3 Usable Security Features

Our participants anticipated a wide range of security measures other than passwords in a crypto wallet in the pre-experiment interview, including pincode (W19), two-factor authentication such as Google Authenticator (W24, W25), account/password recovery, e.g., when losing the initial device (W25), and biometric authentication, such as fingerprints and face recognition (W19, W24), especially in the transaction process where stakes became higher. After the study sessions, all participants felt iWallet was secure. For instance, W25 thought iWallet was safe since *"the secret recovery phrase is encrypted [into the downloaded file] and cannot be read by others."* W22 both downloaded and manually wrote down the secret recovery phrase, and thought it as two-factor authentication. W23 recalled the instructional videos reminded him about password security: *"[It] should be strong, not a human or pet name."*

The security feature of asking users to re-type and check the last 4 characters of the destination address was well received by our blind participants (W21, W22, W25, W26), who could not easily notice different addresses. W21 interpreted the feature as a way to verify who to send money to, which was a common understanding among them. W25 added that in regular apps, when people made transactions, there would be prompts like *"are you sure it's the right person?"* W22, who reconfirmed the last 4 characters multiple times, thought it as a very useful option. However, one participant (W25) noted that it could be hard to catch the last 4 characters with a screen reader and retype them in a single pass.

7 Discussion

In our study, we utilized a competitive analysis of existing wallets, semi-structured interviews, and task-based experiments to identify major accessibility issues experienced by users which led to both learnability and security challenges for blind users. We further implemented and evaluated accessibility-centered design solutions for crypto wallets. Our findings shed light on addressing the accessibility, security, and learnability challenges faced by blind users when using crypto wallets.

In this section, we reflect on accessibility issues in the emerging application domain of crypto wallets and how we alleviated some of these challenges through an iterative design approach. In addition, we discuss education, usable design, and security interventions for improving accessibility. Our goal is to contribute to empirical knowledge in accessibility issues associated with crypto wallets; particularly the intersection of accessibility and security for blind users [12, 46].

7.1 Accessibility Helps Usable Security

Accessibility is of utmost importance for security for blind users. Many user-facing security concerns in the crypto space stem from user misconceptions [40, 41]. Inaccessible wallets leave blind users especially prone to these misconceptions, and in turn, vulnerable to security breaches. We extend prior work [46] by identifying several instances where security information and consequences are not effectively communicated to users via design and assistive technology in the context of crypto wallets. Participants in our study identified ill-fitting security and crypto concepts on MetaMask, making it difficult for them to internalize the consequences and make informed decisions during tasks. For example, seed phrase was not conceptually perceptible for blind users to understand the length of security exploitation if they did not save it securely. In addition, MetaMask provided little audio guidance for web elements, misleading screen reader users during use.

To address these design issues, iWallet incorporated redesigned features that better served blind users in transitioning towards more beneficial states of security awareness. For example, we implement a feature prompting users to confirm/retype the receiving address during transactions, which helps blind users ensure they are sending crypto to the correct addresses, and potentially helps them avoid "clipboard hijacker" which replaces crypto wallet addresses with lookalikes [49]. This design was especially praised by our blind participants since they could not visually differentiate between different wallet addresses like sighted users and were thus more prone to accidentally sending assets to the wrong addresses. MetaMask did not require address confirmation, and several blind users sent assets to their own addresses instead of the one provided by the research team. In short, by improving accessibility, we also improved security.

Another key challenge of crypto wallets for our novice user participants was the lack of fundamental knowledge about cryptocurrencies and wallets [21]. Therefore, we incorporated educational materials about crypto concepts and security into our wallet redesign. We emphasize the importance of learning these critical yet hard-to-grasp concepts by closely collaborating with blind users through the iterative design process. Technical concepts are difficult for novice users to grasp. Accessible designs such as textual explanations about these

concepts should be included which blind users can access and benefit from. These accessible educational materials can contribute to the usable security for blind crypto wallet users.

7.2 Improving Crypto Wallet Accessibility

Though accessibility issues in crypto wallets are well known [27], little effort has been devoted to understanding and improving them. In our user testing of MetaMask, we found that blind users encountered numerous unlabeled buttons, rendering them unable to complete critical steps, including revealing secret recovery phrases, entering transaction amounts, etc. Some participants skipped the secret recovery phrase backup process during onboarding due to the frustrating accessibility, which could potentially lead to severe security risks. Our redesign and evaluation indicated that adhering to accessibility standards and best practices such as WCAG [8] and ARIA [2] could address some critical accessibility issues (e.g., unlabeled buttons) and foster greater adoption of crypto wallets. Future design of crypto wallets should explicitly consider accessibility and implement accessibility best practices.

Moreover, blind users have their unique challenges and needs than sighted users. Visual cues may be quickly grasped by sighted users. For example, after seeing the wallet address string, they immediately knew it was the wallet address without explicit explanations. However, blind users could hardly use such visual evidence to infer unfamiliar concepts. It is also harder for blind users to infer the status of an operation, e.g., if a transaction is confirmed, while sighted users find less difficulty spotting content change on the page, e.g., the change of ETH balance. To improve accessibility for blind users, it is essential to provide more explicit explanations and notifications that can boost their confidence in using crypto wallets. While our notifications in form of popup windows are not accessible enough for blind users, we recommend providing sound notifications which are clear and distinguishable [19]. For example, different sounds can be used for different types of notifications, such as new transactions or errors.

Similarly, we observed that some simple operations for sighted users were conceptually or practically harder for blind users. For example, confirming the secret recovery phrase by re-arranging the shuffled 12 words into the original order was extremely hard with a screen reader. The blind users had to go back and forth to check each word and select them in MetaMask. Correspondingly, we designed an additional option for users to download and upload the secret recovery phrase seamlessly, which was well received by them, finding it time-saving and convenient. The encryption feature further added to the perceived security of the wallet.

Through our iterative design and evaluation, we are aware that adhering to accessibility standards and guidelines is important, but not enough. Application context is crucial for accessibility development. In financial scenarios, blind users expect more notifications and accessible security measures to allow them to navigate the pages with confidence. Thus crypto developers, who are often sighted, should consider accessibility as a key design requirement for their wallet design. Both accessible content features (e.g., educational videos, screen reader compatible text descriptions) and design features (e.g., gas fee slider, downloadable encrypted secret recovery phrase, receiving address checking) could be helpful for blind users.

7.3 Limitations and Future Work

There are a few limitations of our presented work. First, our user studies and redesigns only focused on MetaMask and thus we can not claim that the set of crypto wallet accessibility issues we found was exhaustive. However, these issues were common in other wallets as shown in our competitive analysis. Second, while our sample size is on par with or even larger than that of many other usable security studies with blind users, a larger sample of blind users with diverse (technical) backgrounds would be useful. Third, our tasks may not fully simulate and reveal users' actual behaviors in real transactions when stakes become higher (e.g., transactions involving their own real crypto assets). Future work could deploy the crypto wallet and study user behavior in practice. Last but not least, there are limitations of our accessibility designs. For example, encrypting the secret recovery phrase with a PIN means that users need to safeguard and recall the PIN, which could also become a target for attackers. Future work could consider addressing these limitations and proposing additional designs, e.g., security mechanisms that completely remove the need for having the secret recovery phrase.

8 Conclusion

We presented an iterative redesign of MetaMask to make it more accessible, usable, and secure for novice, blind users. Building on the perspectives of 23 blind users, one low-vision user, and an additional 20 sighted users (N=44), we uncovered a number of accessibility problems with MetaMask that frustrate blind users and that leave them disproportionately prone to common security risks. Our summative evaluation suggests that our redesign alleviated many of these problems: e.g., we followed best practices for accessible design to allow screen readers to access the user interface elements, presented an option for downloading one's secret recovery phrase in an encrypted file to circumvent the need to manually write it down, and created accessible text- and video-based guides to help users understand crypto concepts. Building on these findings, we proposed design implications for creating a more accessible and secure crypto infrastructure for blind users.

References

[1] Ada, Accessed on 2022. https://www.ada.gov/.

[2] Aria, Accessed on 2022. https://www.w3.org/WAI/standards-guidelines/aria/.

[3] Crypto concept, Accessed on 2022. https://techcrunch.com/.

[4] Crypto gaming, Accessed on 2022. https://beincrypto.com/do-web3-gaming-have-a-user-experience-problem/.

[5] dapp auth web3, Accessed on 2022. https://web3auth.io/docs/overview/web3auth-for-dapps.

[6] nfb, Accessed on 2022. https://nfb.org/images/nfb/publications/bm/bm22/bm2210/bm221012.htm.

[7] Uniswap, Accessed on 2022. https://medium.com/@jaddison01/uniswap-a-ux-case-study-ad2088b7fb3.

[8] Wcag, Accessed on 2022. https://www.w3.org/WAI/standards-guidelines/wcag/.

[9] Webaim, Accessed on 2022. https://webaim.org/resources/contrastchecker/.

[10] xmind, Accessed on 2022. https://www.xmind.net.

[11] Svetlana Abramova, Artemij Voskobojnikov, Konstantin Beznosov, and Rainer Böhme. Bits under the mattress: Understanding different risk perceptions and security behaviors of crypto-asset users. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021. https://informationsecurity.uibk.ac.at/pdfs/CHI2021_Bits_Under_the_Mattress.pdf.

[12] Tousif Ahmed, Roberto Hoyle, Kay Connelly, David Crandall, and Apu Kapadia. Privacy concerns and behaviors of people with visual impairments. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 3523–3532, 2015.

[13] Ulun Akturan and Nuray Tezcan. Mobile banking adoption of the youth market: Perceptions and intentions. *Marketing Intelligence & Planning*, 30(4):444–459, 2012.

[14] Hayder Albayati, Suk Kyoung Kim, and Jae Jeung Rho. Accepting financial transactions using blockchain technology and cryptocurrency: A customer perspective approach. *Technology in Society*, 62:101320, 2020.

[15] Hayder Albayati, Suk Kyoung Kim, and Jae Jeung Rho. A study on the use of cryptocurrency wallets from a user experience perspective. *Human Behavior and Emerging Technologies*, 3(5):720–738, 2021.

[16] Andreas M Antonopoulos and Gavin Wood. *Mastering ethereum: building smart contracts and dapps*. O'reilly Media, 2018.

[17] Mark Blythe, Helen Petrie, and John A Clark. F for fake: four studies on how we fall for phish. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 3469–3478, 2011.

[18] Sheri Byrne-Haber. The accessibility issue more people with disabilities should think about. *Medium*, 2019.

[19] R. F. Cohen, R. Yu, A. Meacham, and J. Skaff. Plumb: displaying graphs to the blind using an active auditory interface. In *Proceedings of the 7th international ACM SIGACCESS conference on Computers and accessibility*, pages 182–183, 2005.

[20] Sauvik Das, Cori Faklaris, Jason I Hong, Laura A Dabbish, et al. The security & privacy acceptance framework (spaf). *Foundations and Trends® in Privacy and Security*, 5(1-2):1–143, 2022.

[21] Andreea-Elena Drăgnoiu, Moritz Platt, Zixin Wang, and Zhixuan Zhou. The more you know: Energy labelling enables more sustainable cryptocurrency investment. *Workshop on Fintech and Decentralized Finance*, 2023.

[22] Shayan Eskandari, David Barrera, Elizabeth Stobert, and Jeremy Clark. A first look at the usability of bitcoin key management. In *arXiv*, 2018. https://arxiv.org/pdf/1802.04351.pdf.

[23] Foundation for Interwallet Operability. Blockchain usability report. 2019. https://fioprotocol.io/wp-content/themes/fio/build/files/blockchain-usability-report-2019.pdf.

[24] Thomas Franck. One in five adults has invested in, traded or used cryptocurrency, nbc news poll shows. *CNBC*, 2022.

[25] Michael Fröhlich, Maurizio Raphael Wagenhaus, Albrecht Schmidt, and Florian Alt. Don't stop me now! exploring challenges of first-time cryptocurrency users. In *Designing Interactive Systems Conference 2021*, pages 138–148, 2021.

[26] M. Fröhlich, F. Waltenberger, L. Trotter, F. Alt, and A. Schmidt. Blockchain and cryptocurrency in human computer interaction: A systematic literature review and research agenda. In *arXiv preprint arXiv:2204.10857*, 2022.

[27] Jon Geater and Marta Piekarska. Blockchain workshop position statement. 2016.

[28] Simin Ghesmati, Walid Fdhila, and Edgar R Weippl. Usability of cryptocurrency wallets providing coinjoin transactions. 2022.

[29] G. R. Gibbs. Thematic coding and categorizing. *Analyzing qualitative data*, 703:38–56, 2007.

[30] L. Glomann, M. Schmid, and N. Kitajewa. Improving the blockchain user experience-an approach to address blockchain mass adoption issues from a human-centred perspective. In *Proceedings of the AHFE 2019 International Conference on Human Factors in Artificial Intelligence and Social Computing*, pages 608–616, 2020.

[31] Sam Goundar and Milind Sathye. Exploring access to financial services by visually impaired people. *Journal of Risk and Financial Management*, 16(2):96, 2023.

[32] Abla L Hamilton and Suzan Revah. The challenges and realities of merging online banks. In *Proceedings of the 2005 conference on Designing for User eXperience*, pages 10–es, 2005.

[33] Uzma Jafar, Mohd Juzaiddin Ab Aziz, and Zarina Shukur. Blockchain for electronic voting system—review and open research challenges. *Sensors*, 21(17):5874, 2021.

[34] H. Jang and S. H. Han. User experience framework for understanding user experience in blockchain services. *International Journal of Human-Computer Studies*, 2022.

[35] F. Kamoun, B. M. Al Mourad, and E. Bataineh. Wcag 1.0 versus wcag 2.0 web accessibility compliance: a case study. In *International Conference on Digital Information Processing, E-Business and Cloud Computing*, pages 94–101, 2013.

[36] Brian Kelly, Liddy Nevile, David Sloan, Sotiris Fanou, Ruth Ellison, and Lisa Herrod. From web accessibility to web adaptability. *Disability and Rehabilitation: Assistive Technology*, 4(4):212–226, 2009.

[37] Kicksecure. Cryptocurrency hardware wallet: Threat model. 2022.

[38] Sven Kiljan, Koen Simoens, Danny De Cock, Marko Van Eekelen, and Harald Vranken. A survey of authentication and communications security in online banking. *ACM Computing Surveys (CSUR)*, 49(4):1–35, 2016.

[39] Joselice Ferreira Lima, Gustavo Miranda Caran, Luiz Fernando R Molinaro, and Daniela Favarro Garrossini. Analysis of accessibility initiatives applied to the web. *International Journal of Web Portals (IJWP)*, 4(4):48–58, 2012.

[40] Alexandra Mai, Katharina Pfeffer, Matthias Gusenbauer, and Edgar Weippl. User mental models of cryptocurrency systems - a grounded theory approach. In *USENIX Symposium on Usable Privacy and Security (SOUPS)*, 2020. https://publications.cispa.saarland/3124/1/Soups_Bitcoin_MM.pdf.

[41] Alexandra Mai, Katharina Pfeffer, Matthias Gusenbauer, Edgar Weippl, and Katharina Krombholz. User mental models of cryptocurrency systems—-a grounded theory approach. 2020.

[42] Daniel A Marquez. *An Attempt at Democratizing Resource Allocation for Social Movements Using Decentralized Autonomous Organizations*. PhD thesis, Massachusetts Institute of Technology, 2021.

[43] Indrani Medhi, Aishwarya Ratan, and Kentaro Toyama. Mobile-banking adoption and usage by low-literate, low-income users in the developing world. In *International-ization, Design and Global Development: Third International Conference, IDGD 2009, Held as Part of HCI International 2009, San Diego, CA, USA, July 19-24, 2009. Proceedings 3*, pages 485–494. Springer, 2009.

[44] David Meijer and Jolien Ubacht. The governance of blockchain systems from an institutional perspective, a matter of trust or control? In *Proceedings of the 19th annual international conference on digital government research: governance in the data age*, pages 1–9, 2018.

[45] Md Moniruzzaman, Farida Chowdhury, and Md S. Ferdous. Examining usability issues in blockchain-based cryptocurrency wallets. In *International Conference on Cyber Security and Computer Science (ICONCS)*, pages 631–643, 2020. https://link.springer.com/chapter/10.1007/978-3-030-52856-0_50.

[46] Daniela Napoli, Khadija Baig, Sana Maqsood, and Sonia Chiasson. " i'm literally just hoping this will work:" obstacles blocking the online security and privacy of users with visual disabilities. In *SOUPS@ USENIX Security Symposium*, pages 263–280, 2021.

[47] Jakob Nielsen. Iterative user-interface design. *Computer*, 26(11):32–41, 1993.

[48] R. Phillips and H. Wilder. Tracing cryptocurrency scams: Clustering replicated advance-fee and phishing websites. In *IEEE International Conference on Blockchain and Cryptocurrency*, pages 1–8, 2020.

[49] Akshay Pillai, Vishal Saraswat, and Arunkumar VR. Smart wallets on blockchain—attacks and their costs. In *Smart City and Informatization: 7th International Conference, iSCI 2019, Guangzhou, China, November 12–15, 2019, Proceedings 7*, pages 649–660. Springer, 2019.

[50] Kaihua Qin, Liyi Zhou, Benjamin Livshits, and Arthur Gervais. Attacking the defi ecosystem with flash loans for fun and profit. In *arXiv*, 2020. https://arxiv.org/pdf/2003.03810.

[51] Olivier Rikken, Marijn Janssen, and Zenlin Kwee. Creating trust in citizen participation through decentralized autonomous citizen participation organizations (dacpos). In *DG. O 2022: The 23rd Annual International Conference on Digital Government Research*, pages 440–442, 2022.

[52] Jeff John Roberts. Ethereum wallet metamask passes 30m users, plans dao and token. *Decrypt*, 2022.

[53] Tanusree Sharma, Zhixuan Zhou, Yun Huang, and Yang Wang. " it's a blessing and a curse": Unpacking creators' practices with non-fungible tokens (nfts) and their communities. *arXiv preprint arXiv:2201.13233*, 2022.

[54] S. Suratkar, M. Shirole, and S. Bhirud. Cryptocurrency wallet: A review. In *4th International Conference on Computer, Communication and Signal Processing*, pages 1–7, 2020.

[55] de Paz-Báñez Teresa and A Manuela. Globalisation: Analysis of european countries.

[56] Sri Aravinda Krishnan Thyagarajan, Adithya Bhat, Bernardo Magri, Daniel Tschudi, and Aniket Kate. Reparo: Publicly verifiable layer to repair blockchains. In *Financial Cryptography and Data Security: 25th International Conference, FC 2021, Virtual Event, March 1–5, 2021, Revised Selected Papers, Part II*, pages 37–56. Springer, 2021.

[57] Artemij Voskobojnikov, Borke Obada-Obieh, Yue Huang, and Konstantin Beznosov. Surviving the cryptojungle: Perception and management of risk among north american cryptocurrency (non)users. In *International Conference on Financial Cryptography and Data Security*, 2020. http://lersse-dl.ece.ubc.ca/record/334/files/voskart_fc20.pdf.

[58] Artemij Voskobojnikov, Oliver Wiese, Masoud Mehrabi Koushki, Volker Roth, and Konstantin Beznosov. The u in crypto stands for usable: An empirical study of user experience with mobile cryptocurrency wallets. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021.

[59] Artemij Voskobojnikov, Oliver Wiese, Masoud Mehrabi Koushki, Volker Roth, and Konstantin Beznosov. The u in crypto stands for usable: An empirical study of user experience with mobile cryptocurrency wallets. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2021.

[60] Qin Wang, Rujia Li, Qi Wang, and Shiping Chen. Non-fungible token (nft): Overview, evaluation, opportunities and challenges. *arXiv preprint arXiv:2105.07447*, 2021.

[61] Brian Wentz, Kailee Tressler, et al. Exploring the accessibility of banking and finance systems for blind users. *First Monday*, 2017.

[62] Tao Zhou. Examining mobile banking user adoption from the perspectives of trust and flow experience. *Information Technology and Management*, 13:27–37, 2012.

[63] Z. Zhou and B. Shen. Toward understanding the use of centralized exchanges for decentralized cryptocurrency. In *arXiv preprint arXiv:2204.08664*, 2022.

A Evaluation and User Review of Wallets

With our goal of improving usability, before initial designs, we got familiar with and systematically evaluated 10 popular crypto wallets on Ethereum, in 3 aspects, i.e., education, security, and accessibility. Since some wallets were developed and used in multiple platforms (i.e., mobile, browser, and desktop), we evaluated them independently, and analyzed the consistency across different platforms. These 10 wallets are: MetaMask (Browser & Mobile), MyEtherWallet (Browser & Mobile), Trust (Mobile), Coinbase Wallet (Mobile), Exodus (Desktop & Mobile), Argent (Mobile), Jaxx Wallet (Desktop & Mobile), DeFi Wallet (Mobile), Lumi (Mobile), and Atomic (Desktop & Mobile). We also analyzed user reviews from various sources such as Chrome Web Store (chrome extension), App Store (iOS), and Google Play Store (Android) to gain additional insights on each wallet.

Evaluation *General Features.* According to our evaluation, we found that the wallets contained similar functions and processes: an onboarding process where users got and were asked to keep their seed phrase, which was the only way to access their wallets, buying/sending/receiving/exchanging crypto assets, and displaying chart information about each asset's market values and trends. Table 2 lists features in different wallets.

Cross-platform wallets allowed users to access their assets on different devices and with different operating systems. However, only half of the wallets evaluated had an alternate platform besides mobile. This could be an accessibility concern for those who wanted to use a crypto wallet but did not have a smart phone.

A notable feature comes from the DeFi mobile wallet, which integrates a status meter for users to check on current gas prices, with estimated transaction times, cryptocurrency values, USD values, and graphics for showing how much traffic is currently on the Ethereum network. This feature may help users make better-informed financial decisions and

Wallet	*Multiple Platform*	*Seed Phrase Security*	*Multi-sig Security*	*Two-factor Auth*	*Education Page*	*Gas/Traffic Meter*	*Dapp Integration*	*Accessibility*
MetaMask	✓	✓	✗	✗	✗	✗	✓	✓
MyEtherWallet	✓	✓	✗	✗	✓	✗	✗	✗
TrustWallet	✗	✓	✗	✗	✓	✗	✓	✗
Coinbase	✓	✓	✗	✗	✗	✗	✓	✓
Exodus	✓	✓	✗	✗	✓	✗	✗	✗
Argent	✗	✗	✓	✗	✗	✗	✓	✗
Jaxx	✓	✓	✗	✗	✗	✗	✓	✗
DeFi	✗	✓	✗	✓	✓	✓	✗	✗
Lumi	✗	✓	✗	✗	✗	✗	✗	✗
Atomic	✓	✓	✗	✗	✓	✗	✗	✗

Table 2: Competitive analysis of existing crypto wallets.

develop a greater awareness of how gas prices work. Our later design of the gas fee slider has been inspired by this feature.

Education. A main usability issue is the lack of educational resources for users. Half of the wallets contained education resources and features, but typically only provided one or two screens. MyEtherWallet (mobile) was the only wallet that had a dedicated resource page, educating users on cryptocurrency, blockchain technology, security/privacy best practices, and the wallet itself. However, these resources were not properly embedded in the browser version of MyEtherWallet: the resources were externally provided, and one must log out of their wallet to access them.

MetaMask (mobile) was the only wallet that provided a tutorial for users, teaching them about different parts of the app and what they could do. However, this tutorial was only accessible as a part of the onboarding process, with no other way of accessing it in the future or daily use. This tutorial was also completely absent from its browser extension counterpart, again displaying cross-platform inconsistency.

Several wallets included tips or popup screens that appeared after clicking a "help" icon, especially for certain screens/features which required additional clarification. But they were not present in other screens where a user may need help or instructions. For example, MetaMask (mobile), Argent, and DeFi only provided instruction in the settings section, specifying their security measures. The remaining screens were designed with the assumption that users understood their content.

Security. Out of the 10 wallets evaluated, nine used a mnemonic/seed phrase for account backup and recovery. The only wallet without a seed phrase mechanism, Argent, alternatively utilized a multi-signature authentication for the same purpose. Specifically, one could choose trusted people as "guardians" to help recover the wallet and approve transactions. Other common security measures in these wallets included daily transaction limits, auto-locking, and 2-factor authentication via Web2 intermediaries such as phone/email or a third-party authenticator.

Some wallets, such as MetaMask chrome extension, does little to guarantee security/accuracy of transactions. Specifically, it did not include a confirmation mechanism to verify receiving address during performing transaction. As a result, users would miss the opportunity to double check, which may lead to sending crypto to a wrong address.

Accessibility. During the wallet accessibility evaluation, we used a contrast checker [9] to check the color contrast of MetaMask as well as Coinbase wallet. In MetaMask, the contrast ratio between foreground color (i.e., text) and background color was 4.27 : 1, which did not meet the standard of WCAG 2.0 level AA, which is a standard to assess web accessibility compliance [35].

User Reviews We analyzed users reviews to gain additional insights on each wallet, and identify features/functions users wanted. To have an abundant collection, we collected reviews for each wallet from Apple App Store (for iOS version), Google Play Store (for Android version), and browser stores of Chrome and FireFox (for browser extension version). From the thematic analysis of the user reviews, we similarly found three main areas of improvement, i.e., education, security, and accessibility. Here we present representative issues encountered by users, as well as design suggestions and preferences provided by them.

Education - MetaMask [Firefox Browser Add-on] Users frequently indicated the need for educational resources and instructions during the onboarding process of crypto wallets. One such example was from a MetaMask user who complained about the lack of (jargon-free) instructions in MetaMask: *"I'm a newby with little experience. The instructions from MetaMask are confusing and often lead to screens where there are no direct instructions on what to do next. MetaMask seems to lack personnel who can write instructions in a clear (without jargon) style." - Firefox user 16749602, 1 star (03/12/2021)"*

Security - MetaMask [Google Play Store] The limited usability and security afforded by seed phrase is repeatedly complained about. Some users mentioned the current process of dealing with the seed phrase was time consuming, and less usable than the traditional security measures they usually used

in Web2, such as Google Authenticator: *"I think there could be more security options, like google authenticator, email verification! A friend of mine was hacked last week, by some sort they got possession of his 12 word seed phrase and boom 5k lost!" - Pedro Santos, 4 stars (05/07/2021)*

Accessibility - MyEtherWallet [App Store] Yet another important area of improvement is accessibility for engaging a broader audience to the crypto space. Regarding accessibility, many users encountered unlabled buttons and input fields while using crypto wallets, e.g., in the following review, *"Buttons need to be labeled for TalkBack users who are blind. I'm unable to create or sign in to the app, because a fair number of the buttons are not labeled so I can't tell if I'm entering in something correctly or not let alone if it is correct and heading the correct or wrong button." - chuck winstead, 1 star (03/23/2021)*

The above analysis of wallet features and user reviews has heavily impacted our redesign process.

B Participant Demographics

We strove to recruit a diverse set of participants for our user evaluations, in terms of age, gender, profession, and vision ability. By so doing, we aimed to identify unique needs of potential crypto users, especially those who were blind, and inform inclusive designs for our crypto wallet. Most participants did not have prior experiences with cryptocurrencies before the study. Only six out of 44 had some prior experience with centralized exchanges such as Binance and Coinbase. More demographic details of our participants are summarized in Table 3.

C Pilot Study Results of iWallet (V1)

The pilot results of iWallet (V1) showed that the educational materials helped users understand cryptocurrency and wallet concepts and terms. Most participants deemed the videos useful and were willing to watch them. However, some skipped videos due to their habit (e.g., W1), perceived uselessness (e.g., W10, *"I skip everything but can still get it set up"*), and preferring text over videos (e.g., W15, *"I'm a text person"*). W5 suggested making the videos mandatory, which could help novice users learn the important basics. W3 further suggested embedding a video to educate users on crypto-related laws. On average, our participants got 3.7 knowledge questions correctly before using the wallet, and that average increased to 4.4 after using the wallet, showing an 19% improvement. In addition, nearly all blind users expressed that they preferred text instructions over video instructions since text were more accessible and saved time.

The security feature of confirming the receiving address was positively received by the participants. All of them expressed that the design made them more confident during the transaction process.

Our accessibility improvements led to less time for task completion for blind users. It took them 40 minutes on average to finish all the tasks, while in the MetaMask evaluation, the blind users spent an average of 47.9 minutes on the tasks. However, we identified several accessibility challenges in our initial redesign. For example, our blind participants suggested putting explicit text explanations, such as "wallet address," next to the elements, to help them more easily understand crypto concepts. The button labeling was not intuitive and informative enough for many blind users, e.g., when they clicked the Next button: *"Next to what? More information is needed."* Our blind participants further expressed that videos should be complemented by text summarizing their content, which was a favored medium for them to save time. The option of downloading the encrypted version of the secret recovery phrase was deemed helpful for blind users. However, we did not prioritize the option of uploading the secret recovery phrase in the account importing process. W11 expected the upload option on the left in the account importing page but only saw the "typing the phrase" option. She failed to import her wallet since she typed in the encrypted phrase which was downloaded during the onboarding instead of the plain text phrase. As a result of this inconsistent design, six out of eight blind participants were unable to import their wallets successfully. This highlighted the importance of consistency of function layout, especially for screen reader users who had more difficulty navigating. The accessibility glitches explained our relatively low average score (65 out of 100) in the SUS survey.

ID	Platform	Country	Gender	Age Group	Occupation	Crypto Experience	Visual Impairments
M1	MetaMask	Germany	Female	25-34	PhD Student	N	No
M2	MetaMask	China	Female	18-25	Master Student	Y	No
M3	MetaMask	China	Female	25-34	Master Student	Y	No
M4	MetaMask	Sweden	Female	18-25	Master Student	N	No
M5	MetaMask	US	Female	18-25	PhD Student	N	No
M6	MetaMask	China	Male	18-25	Data Scientist	N	No
M7	MetaMask	US	Male	18-25	Self-employed	N	No
M8	MetaMask	Netherlands	Male	25-34	Vehicle Engineer	N	No
M9	MetaMask	India	Male	25-34	PhD Student	Y	No
M10	MetaMask	Switzerland	Male	18-25	Administration	N	No
M11	MetaMask	US	Male	45-54	Accessibility Expert	Y	Blind
M12	MetaMask	US	Female	35-44	Logistics Management Specialist	N	Blind
M13	MetaMask	US	Female	45-54	Contractor	N	Blind
M14	MetaMask	US	Male	25-34	PhD Student	N	Blind
M15	MetaMask	US	Agender	35-44	Freelance Artist	N	Low-vision
M16	MetaMask	US	Male	25-34	Master Student	N	Blind
M17	MetaMask	US	Male	45-54	Financial Consultant	N	Blind
M18	MetaMask	US	Female	25-34	Self-taught Student	N	Blind
W1	iWallet(V1)	US	Male	18-25	Master Student	N	No
W2	iWallet(V1)	US	Female	25-34	PhD Student	N	No
W3	iWallet(V1)	Spain	Female	18-25	Translator	N	No
W4	iWallet(V1)	US	Male	18-25	Store Worker	N	No
W5	iWallet(V1)	Netherlands	Male	18-25	Journalist	N	No
W6	iWallet(V1)	Nigeria	Female	18-25	Undergrad Student	N	No
W7	iWallet(V1)	China	Female	18-25	Accountant	N	No
W8	iWallet(V1)	Hong Kong	Male	18-25	Master Student	Y	No
W9	iWallet(V1)	China	Female	18-25	Project Manager	N	No
W10	iWallet(V1)	US	Male	25-34	PhD Student	N	No
W11	iWallet(V1)	US	Female	25-34	Unemployed	N	Blind
W12	iWallet(V1)	US	Female	55-64	Accessibility Specialist	N	Blind
W13	iWallet(V1)	US	Male	25-34	Accessibility Evangelist	N	Blind
W14	iWallet(V1)	Italy	Female	35-44	PhD Student	N	Blind
W15	iWallet(V1)	US	Male	18-25	Undergrad Student	N	Blind
W16	iWallet(V1)	Bahrain	Male	25-34	Administration	N	Blind
W17	iWallet(V1)	UK	Male	35-44	Open-source Developer	N	Blind
W18	iWallet(V1)	Canada	Male	25-34	Developer	N	Blind
W19	iWallet(V2)	US	Female	35-44	Accessibility Consultant	N	Blind
W20	iWallet(V2)	US	Male	18-25	Lab Scientist	N	Blind
W21	iWallet(V2)	US	Male	18-25	Therapist	N	Blind
W22	iWallet(V2)	India	Female	35-44	Bank Manager	N	Blind
W23	iWallet(V2)	India	Male	25-34	Partnership Specialist	N	Blind
W24	iWallet(V2)	UAE	Female	18-25	High School Graduate	N	Blind
W25	iWallet(V2)	US	Female	45-54	College Student	N	Blind
W26	iWallet(V2)	US	Female	55-64	Self-employed	Y	Blind

Table 3: Demographic information about the participants. M1-M18 tested MetaMask. W1-W18 tested iWallet (V1). W19-W26 tested iWallet (V2).

Youth understandings of online privacy and security: A dyadic study of children and their parents

Olivia Williams, *University of Maryland*[1], *College Park, MD, USA*
Yee-Yin Choong, *National Institute of Standards and Technology, Gaithersburg, MD, USA*
Kerrianne Buchanan, *National Institute of Standards and Technology, Gaithersburg, MD, USA*

Abstract

With youth increasingly accessing and using the internet, it is important to understand what they know about online privacy and security (OPS), and from where they gain this knowledge in order to best support their learning and online practices. Currently, the field of literature surrounding such youth understandings has gaps in depth and breadth that we aimed to address in this study. We conducted semi-structured interviews with 40 youth/parent dyads with youth in 3rd-12th grades in the United States to understand more about what youth know about OPS and how their parents attempt to influence this knowledge. We found that youth of all ages in the study could provide at least basic descriptions of both online privacy and online security and could give relevant examples of good and bad OPS choices. We also found that parents took a variety of approaches to influencing youth understandings and behavior, with most of those approaches relying on device monitoring and limiting use. However, parents who attempted to influence their children's knowledge through conversations had children who demonstrated the most nuanced understandings. Our findings offer promising suggestions for parents, technology providers, and future research.

1. Introduction

Children and teenagers under 18 (hereafter referred as "youth") utilize technology more and at younger ages than ever before [1], and are often "digital by default" [52] with digital footprints that begin before birth [53]. In 2019, 95% of 3–18-year-olds in the United States had home internet access [56]. With this access, youth of all ages participate in a variety of activities online—including gaming, researching, social media, emailing, and streaming entertainment ([1][28])—all of which involve elements of online privacy, security, and personal data management. As a result of this ongoing use, youth's descriptions and understandings of online privacy and security (OPS)[2] are constantly in flux as they learn how to protect themselves and be responsible in an ever-evolving online context ([22][60]).

To know how to best support youth's ever-developing OPS knowledge, we need to know more about the influences and intricacies of their current understandings, as well as the people and places—including parents, family members, schools, friends, and technology itself—that influence those understandings.

Quayyum and colleagues [44] reviewed a decade of youth cybersecurity awareness literature from 2011-2020 and concluded that "although cybersecurity awareness research for children has received significant attention from researchers, there remain gaps," particularly in evaluating youth's awareness [44]. The purpose of our study was to address this gap in order to learn more about what youth know about OPS. In addition, we wanted to understand how parents understand and attempt to influence youth's OPS knowledge and practices given the important role they play in youth's lives and access to technology. To achieve these purposes, we interviewed 40 youth/parent dyads with youth in 3rd-12th grades to answer three research questions:

1. What are youth's descriptions of online privacy and online security, and how do they understand these terms?
2. How do parents view the role of online privacy and online security in their children's lives?
3. How, if at all, do parents influence children's online privacy and online security understandings?

Our qualitative investigation of these questions is unique in two ways. First, the design of our study was distinct from its peer research in both its dyadic structure and in the broad age range of youth participants (3rd-12th grade). This design and population allowed us to compare youth and parent

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.

USENIX Symposium on Usable Privacy and Security (SOUPS) 2023.
August 6 -- 8, 2023, Anaheim, CA, USA.

[1] This material is based on work supported by the UMD and NIST Professional Research Experience Program (PREP) under Award Number 70NANB18H165.

[2] We use OPS as an acronym for brevity when we are talking broadly about the focus of the study *or* discussing parental involvement in aspects of youth online activity that influence both online privacy and security.

understandings surrounding OPS across and within dyads and grade bands to look for interrelationships that cannot be studied using other research designs. Second, much of the extant literature surrounding online security knowledge, specifically, examines youth's knowledge after participating in some kind of learning experience like a cybersecurity game or summer camp (e.g., [24][50]). These studies are valuable in that they evaluate contexts and supports that help youth learn about OPS, but they also precludes the ability to know what and how children understand OPS in their day to day lives either before or without such targeted learning experiences. Our study, by contrast, explores youth's knowledge without any kind of OPS knowledge intervention in order to better examine what youth authentically know about OPS.

For the purposes of this study, we acknowledge that both "online privacy" and "online security" are broad, complex terms for which descriptions depend on audience and context. There is no commonly and widely used description of either term, which has been recently acknowledged by the field (e.g., [20][22][42]). Accordingly, in this paper we explore extant research involving youth knowledge of each term separately ("online privacy" and "online security"), and report out findings regarding our study participants' understandings of each term separately. However, we purposefully did not provide our own definitions of these terms given the study's overarching purpose of understanding how our participants describe and understand the terms through their own words and examples.

2. Related Work

2.1. Youth OPS Understandings

2.1.1 How Youth Understand Online Privacy

Youth's "needs, opinions, experiences, and attitudes towards privacy and data protection are the least researched so far"[37], due largely to the fact that many adults believe youth, especially young children, are too young to understand or care about online privacy [29]. That belief, however, is inaccurate. The research reviewed for this study agrees that youth as young as six have some knowledge and care about the basic idea of online privacy ([37][67]). These studies also reveal that although youth find online privacy important and have basic ideas about why it matters, these understandings are not nuanced.

Younger youth especially (up to age 11) have been found to value their privacy without fully understanding what it means to be private online [65], and to have flawed reasoning behind their understandings [5]. Youth in this age group have also been found to dislike the idea of their personal information being shared with strangers online, but do not always know how to prevent this from happening [44]. These gaps in youth's understandings are partially attributable to developmental processes: the concept of "online privacy" contains both tangible and abstract aspects that are complex, varied, and constantly changing ([8][18]). This can make it difficult for people of any age to learn about and exercise good online privacy behaviors, but especially youth, who do not begin processing abstract concepts until around age 12 [45]. Even as youth move into their teenage years, navigating abstract concepts like "privacy" and "security" takes time, and can be difficult to translate into online practice [45].

Further, online privacy can be broadly categorized into three "levels"—the interpersonal, the institutional (i.e., government organizations), and the commercial—that are all unique and need to be understood differently [32]. Because youth's developing online privacy understandings are an extension of their knowledge of privacy in the off-line world, they often have a strong sense of interpersonal online privacy (i.e., avoiding "stranger danger"), but have much less institutional or commercial privacy knowledge [52]. This is problematic, because the moment youth exist and interact online, their information and data are being collected. However, youth often have no idea what that means or what (if anything) they should do about it ([33][50]). This results in youth who are "cautious about strangers…[but] have not yet received knowledge about how corporate forces can use their data" [8]. This sentiment was echoed across multiple studies (e.g., [12][50]), with Milkaite and colleagues noting "when it came to their participants' [83 9-12-year-old] knowledge of data protection rights and of more detailed data processing actions, the purposes of data collection, sharing and general use in commercial and institutional contexts, children's understanding was much more limited" [37].

Finally, some studies show that youth—and especially older youth—view elements of online privacy as negotiable and choice-based, which contributes to something called the "privacy paradox" [23]. The "privacy paradox" is the notion that privacy knowledge does not always translate into privacy-protective strategies. For example, in a study of 366 4th-6th graders, this discrepancy between what youth know about privacy and if or how they put that knowledge into practice was common, and was most striking in the oldest (6th grade) youth in the study [15]. Similarly, a survey of 805 9-17-year-olds in Taiwan revealed that "performing privacy protective practices did not simply lead to fewer privacy-precarious practices" [10], suggesting that youth who had knowledge of online privacy took preventative measures while also maintaining questionable practices, or intentionally did not practice making choices that align with their knowledge of online privacy ([7][44]). It is important to note that the privacy paradox has been critiqued for its inability to sufficiently address the pervasiveness of technology in everyday life [60], and these critiques are in line with other calls to better address how the field needs to work on redefining ideas like "privacy" to capture new contexts like digital platforms [3]. Many such critiques include the online technologies and

platforms most used by youth, making the critiques particularly important topics to consider in the youth user context.

2.1.2 How Youth Understand Online Security

A majority of studies exploring youth understandings surrounding online security focus on the impact of online cybersecurity games or interventions on youth knowledge (e.g., [11][41]). The result is that these studies do little to help explain what youth actually know without receiving targeted training first, or what they take away from learning about online security outside the minutes immediately following some sort of targeted instruction. It is likely that some of the same challenges with abstractness and developmental thinking that can make online privacy challenging to learn also apply to the process of learning about online security, but the topic is less well studied and understood.

In one of the few investigations seeking to broadly understand youth's cybersecurity awareness, only 19% of the 2,214 8-12-year-olds and 32% of the 13-17-year-olds surveyed in New Zealand recognized seven common cybersecurity terms [57]. Of those who did recognize terms, most of the awareness surrounded more fundamental ideas like firewalls and antivirus software with very few youth—only one of the 444 youth in the 8-12-year-old group—having an awareness of terms like "phishing" and "tracker" [57]. Other cybersecurity awareness surveys were conducted in Malaysia [68] and Turkey [63] with similar results: youth were found to have very basic levels of cybersecurity knowledge and awareness, and rarely took measures to increase their cybersecurity.

Most of the other available literature addressing youth's preexisting online security knowledge uses passwords as a vehicle to gauge this understanding, and this password-related literature is also reflective of the knowledge vs. practice paradox. For example, Theofanos and colleagues [56] found that in 8-18-year-olds, older youth had more password knowledge, but were also more likely to report using poor password practices like sharing passwords with friends or reusing passwords across multiple sites [56].

2.2. Framing Youth Knowledge and Behavior Through a Social Learning Lens

In this study, we use social learning theory [4] to frame our understanding of youth OPS knowledge and parents' potential influence on that knowledge. Social learning theory suggests that most human behavior is learned observationally through modeling and from one's surroundings; people learn from seeing or being taught something, trying it on their own, and then evaluating the results [4]. Through this lens, to better understand youth's OPS knowledge and behavior, we must better understand their contextual influences—such as parents, family members, friends, teachers, and technology itself—as well as what motivates youth to retain and actually use OPS best practices. In this study, we chose to specifically examine the contextual influence of parents because of how prevalent and influential parent relationships are in children's lives. A social learning framework led us to focus our data collection and analysis on how youth described and explained OPS and how parents described their roles in their children's OPS knowledge development to examine possible connections between the two.

2.3. Parental Influence

In terms of contextual influences on youth's OPS knowledge, parents are a natural point of inquiry given their central role in youth's lives. Especially up until around age 11, youth rely on their parents for support with OPS choices and tend to seek out and accept parental oversight and support [33]. What extant literature otherwise knows about parents' influence on youth OPS understandings, however, is complicated. For example, Manotipya and Ghazinour surveyed 1,300 parents from 51 countries and found that parents generally feel that they have some awareness of their children's online privacy practices, but that parents also often pose a threat to their children's privacy by oversharing information online ([19][34]).

Device monitoring tends to be a common practice for parents to influence their children's OPS. In an interview study about child internet use and protection strategies with 14 families, 18 protection strategies were found, 17 of which were physical or technical controls like restricting access, configuring privacy settings, and restricting access as punishment [64]. Despite its widespread use, device monitoring may only be effective with younger youth. In a study of 1,700 4^{th}-6^{th} grade students' internet use and supervision, about half of the students reported being supervised when using the internet at home. Those youth who reported some level of parental oversight were more likely to practice privacy protective behaviors [59]. A separate survey of 746 12-18-year-old youth, however, told a different story. Unlike their 4^{th}-6^{th} grade counterparts, the teenagers surveyed by Shin and Kang [48] who experienced device monitoring and use rules did not demonstrate more privacy-protective behaviors. This reflects the conclusions by other scholars that teenage youth do not want to be monitored by parents as much. It also aligns with the demonstrated youth understandings of privacy and security as being choice-based.

Extant literature on parental influence does suggest that conversation and communication are also important ways that parents influence their children's knowledge and behavior, with multiple studies concluding that "internet parenting is best achieved through an open communication style and through making connections with children" ([46][48][53]). Unfortunately, parents experience challenges with communicating with their children about OPS topics. Some parents feel that their children are too young to understand or exercise protective OPS behaviors, and admit that cybersecurity conversations at home are not common ([27][40][64][67]). Other

parents have noted feeling like their own understandings are not strong enough to know how to protect their children ([15][32][66]). In these instances, especially with older youth, it may be difficult for parents to make meaningful contributions to their children's knowledge [66].

In summary, existing research on youth online privacy and security knowledge suggests that they have some understanding of these terms, but may not always put this knowledge into practice. In this literature, however, previous studies have focused on either online privacy or online security separately without differentiating between the terms. Additionally, past studies tend to examine a narrow age range of youth, and/or are tied to specific knowledge interventions. Our study aims to contribute to this field by investigating privacy and security knowledge in tandem (and thus exploring if youth can differentiate between the two concepts) and studying a broader age range of youth in order to compare knowledge across grade bands. We also seek to better understand youth knowledge *in situ* as opposed to in response to a learning task. Further, we aim to investigate youth knowledge alongside parental knowledge and understanding because of the important role that parents can play in shaping youth's knowledge and behavior development.

3. Methods

To answer this study's research questions about what children know about OPS and how parents attempt to influence that knowledge, we conducted a qualitative study consisting of pre-interview questionnaires and semi-structured interviews with 40 youth/parent dyads in spring 2021.

3.1. Recruitment and Participants

This study was approved by the Institutional Review Board (IRB) of the National Institute of Standards and Technology. Parent/child dyads for this study were recruited by a contracting research firm that used a preexisting user database; eligible parents self-elected themselves and their child for participation. A total of 40 youth/parent dyads from across the United States participated. These dyads included 4 youth from each grade from 3rd-12th grades and one of their parents, resulting in 12 elementary school (ES; 3rd-5th grades, 8 to 11 years old), 12 middle school (MS; 6th-8th grades, 12 to 14 years old), 16 high school (HS; 9th-12 grades, 15 to 18 years old) participants, and 40 parents. Demographic information for each dyad can be found in the table in Appendix A.

3.2. Instruments

Data were collected using a pre-interview questionnaire and a semi-structured interview. The two instruments were designed to be mutually inclusive; the questionnaires collected demographic data and participants' basic descriptions and positions about online privacy and security and served as a pre-thinking exercise for participants for the interview, and the interviews allowed participants to expand upon and discuss their answers from the questionnaire with thoughts, examples, and personal narratives. The questionnaire language was scaffolded to suit participants' age and role, resulting in three different versions: one for youth in grades 3-5, one for youth in grades 6-12, and one for parents. All three questionnaires consisted of content sections with demographic questions, general technology use questions, OPS knowledge questions, and three online risk questions. The parent questionnaire was six questions longer because parents were asked about both themselves and their children.

The semi-structured interview protocols were also scaffolded to suit participants' ages and roles [1]. Youth participants were asked 11 anchor questions about their knowledge of and behavior surrounding online privacy, security, and risk. Parent participants were asked 9 anchor questions about both their own knowledge of online privacy, security, and risk, as well as how they view their child's knowledge and behavior surrounding these ideas.

Two members of the research team—one quantitative expert and one qualitative expert—created an initial draft of the data collection tools using the research questions and extant literature as a guide. From there, the content and quality of both tools were refined over four iterative steps: (1) review by a survey expert, (2) review by research colleagues and four K-12 teachers, (3) cognitive interviews with three youth (one elementary, one middle, and one high schooler) [7], and (4) pilot interviews with three youth/parent dyads [55]. After each step in this process, the data collection tools were refined based on feedback and pilot participant responses. The study instruments are included in Appendix B.

3.3. Procedure

All data collection occurred remotely over Zoom[3] and was audio-recorded for transcription. The youth/parent dyads signed informed consent and assent forms (for youth older than 12) and were briefed about the study together.

Following the study overview and verbal consent/assent process, parent and youth participants were interviewed separately in order to afford both parties—but particularly youth participants—the privacy needed to answer potentially sensitive questions about their online activities as openly and honestly as possible. All parents and youth were given the option to have youth participants interviewed with their parent in the room if they were more comfortable with this option. One

[3] Any mention of commercial products or reference to commercial organizations is for information only; it does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the products mentioned are necessarily the best available for the purpose.

youth participant and one parent participant selected this option, and the other 38 dyads were interviewed separately.

Each of the 40 data collection Zoom calls were scheduled for 90 minutes, and the first author conducted all 80 interviews for consistency. Participants were compensated for their time with cash gift cards: parents received $75 and youth received $25. Any personal identifiable information (such as name and location) unintentionally revealed during the interview was properly redacted and removed from the data. Each participant was assigned a unique alphanumeric identifier. The data collection process yielded 80 complete pre-interview questionnaires and 546 pages of single-spaced interview transcripts.

3.4. Data Analysis

The qualitative data analysis for this study proceeded across two cycles and was guided by methods outlined by Johnny Saldaña [47]. Cycle one contained both inductive and deductive coding resulting in 84 first-cycle codes. This initial code deck was used by the first and third author to code a random selection of nine full dyad transcripts using Nvivo coding software. The full research team then met to discuss and refine the code deck. This process was repeated three more times with different samples of three dyad transcripts in order to refine the code deck. The first and third author then used the fourth revision of the code deck to run an interrater reliability (IRR) agreement statistic, which returned a Cohen's Kappa (k) value of .74 indicating substantial interrater agreement [36]. All coding discrepancies were resolved through discussion with the full research team. Once the IRR statistic was calculated, the first and third author completed a final first-round coding pass of all 40 dyad transcripts.

After all first-round coding was complete, the research team read through the coding results individually, then met to discuss patterns and themes for each research question from a dyadic perspective. To do this, we used the constant comparative method as outlined in Boejie [6] and Williams-Reade and colleagues [61] as mentor processes for how to compare and contrast codes across different participants and dyads seeking patterns, similarities, and differences which could then be categorized and conceptualized. In doing so, we followed these steps: (1) comparison of data and coding within a single participant's interview, (2) comparison between interviews within the same group (all youth and then all parents), (3) comparison between different groups (all youth with all parents), (4) comparison in pairs at the dyad level (individual youth with their parent), and (5) comparison across all dyads. Finally, the first author performed a second cycle theming of the resulting data using the research questions as a frame. A table demonstrating an example of the coding process can be found in Appendix C.

4. Results

The results of this study are evidenced with direct quotes from participants and cited with an alphanumeric identifier. In the identifiers, the "Y" or "P" indicates "youth" or "parent," the number is the dyad code, and the ES/MS/HS indicates whether the youth participant of that dyad was an elementary (3rd-5th), middle (6th-8th), or high school (9th-12th) student.

4.1. RQ1: What do youth know about OPS?

4.1.1 Youth online privacy descriptions

Youth across all grade bands in this study described online privacy as how one protects personal and important information, and often did so in interpersonal terms. In these descriptions, "information" primarily meant personal details such as full name, location, age, passwords, and financial information, and the goal was to keep it from being accessed by strangers, hackers, and other unwanted third parties. Youth described online privacy as a way to "have your independence" (Y03MS), "be safe" (Y06ES), and keep someone from "knowing your own business without you telling them" (Y22HS). The 40 youth also unanimously agreed that online privacy is important.

MS and HS youth also spoke about their online information privacy agentively, positioning it as something over which they had some control. For example, when Y22HS was explaining what he meant by preventing people from "knowing your own business without you telling them," he clarified that "a lot of [sketchy websites or skilled hackers] probably ask for credit card information or an email address or a phone number…the worst thing people can do is to give out information that's not necessarily needed." In this explanation, he positioned the online user as having the choice to either tell/give or not tell/give their information. Through such choices, older youth position online information privacy not as something simply afforded to a person, but instead as an idea that is always under construction and dependent upon an ongoing series of choices.

There was also a recognition across age groups that online privacy is contextual, and that within certain contexts a person can choose how much privacy they want to have. Games and social media served as important examples of this point; for instance, Y21MS explained that she chose to have a private TikTok account because "if I posted a video, I wouldn't want it blowing up to the point where it has a million [views]…it would be overwhelming," but that public accounts are the right choice for some people. Of the youth that discussed social media, all opted for private pages on platforms like Instagram and Pinterest, or stated a preference for apps like Snapchat that require a user to add "friends" before those friends can view shared content.

Finally, youth descriptions and examples of online privacy featured trust and feelings about security as central components for making good online privacy choices. Having a sense for which people and which websites are trustworthy—and, more frequently, which ones are *un*trustworthy—emerged as a way that youth believed they could keep themselves and their information private. For example, Y12ES advised that "no one really that you don't trust should know your private information." The most cited untrustworthy entities were strangers (writ large), hackers, and advertisements or pop-ups.

In terms of *why* they felt online privacy was important, responses overwhelmingly included the consequences of "being hacked and having your information stolen" (Y02MS), having someone "get into your bank account and take your money" (Y10ES), "identity theft" (Y22HS), and having sensitive information like "photos get(ting) leaked, and then it's extremely hard to get those photos off the web…that can affect your online and personal life" (Y27HS). HS youth were more likely to only cite virtual consequences of poor privacy choices like data theft and hacking by "the people that are good with computers" (Y39HS), while ES and MS youth were also frequently concerned about in-person consequences like kidnapping (Y18ES) and people who "could potentially steal from you and come over and rob [you]" (Y33MS).

4.1.2 Youth online security descriptions

Many youths described the online security by giving examples of choices that can be made to either increase or decrease security. Specifically, youth overwhelmingly mentioned good password behavior like making sure "all my passwords aren't the same" (Y21MS), setting "strong passwords" (Y14ES), and not "shar(ing) my passwords and stuff" (Y34HS), as well as broader device and browsing choices like using "a secure network" (Y04HS) and only clicking on "secure websites" (Y03MS). MS and HS youth also noted using certain technologies—like virtual private networks (VPN) and firewalls—to help maintain their online security.

To determine which websites were secure, youth sometimes cited concrete evidence like looking "in the left corner where it has the website link there's usually a green lock" (Y12ES), but also sometimes mentioned relying on simply "feel[ing] like I'm set to just know that I made the right choice" (Y32MS). Some youth also described feeling confident in their online security behavior because they had not (yet) experienced any negative consequences, like Y27(HS) who explained that she knew she was secure online because she "[hadn't] had my information leaked, [or] had any photos leaked."

Across grade bands, online security was described as a way to help ensure online privacy and protect against outside threats, specifically hackers and viruses. There was, however, a grade band difference in how much immediate, personal control youth felt that they had over their online security: youth in ES and lower MS were more likely to rely on parents and/or security software for security, while their upper MS and HS counterparts relied more on themselves and their ongoing choices. For example, Y09ES noted that before downloading apps "I always ask my dad first to see if maybe I could accidentally download a virus or something," While Y34HS reported feeling secure because he used "safe apps and…won't share my passwords."

4.1.3 Youth OPS understandings

While youth's OPS understandings in this study shared many characteristics, there was a difference between youth's privacy and security knowledge, with youth providing more extended and detailed descriptions of *online privacy*. Youth across grade bands were more likely to say things like "I know more with privacy than security" (Y31MS), or to have less depth in their security knowledge, like participant Y10ES who knew "you could add extra security to your device," but could not give an example of how. That being said, most of the youth *were* able to at least differentiate between "online privacy" and "online security," often using the idea of "privacy" in their descriptions of "security," but rarely the other way around. This suggests a specific (rather than random or conflated) understanding about the relationship between the two terms: that good security choices help ensure "that other people are not doing things that could potentially harm you or your privacy" (Y02MS).

Further, outsiders were cited as the biggest threat to both online privacy and online security. However, the nature of the threat was described slightly differently across the terms. With online privacy, youth described the threat as losing anonymity and, along with it, security, while with online security, youth described the threat as having information stolen. This understanding of threat seemed to also translate into an understanding of the role of agency.

The 40 youth in this study described being able to make good and bad choices that could either increase or decrease both their OPS. However, when it came to privacy, these choices were more often described as optional, ongoing, and existing on a spectrum, whereas with security the choices were described as more necessary, one-time in nature, and clear cut. For example, participant Y37HS referred to social media to discuss both her privacy and security choices, but in different ways. When talking about privacy, she described "only letting people that you know and that you are comfortable with follow you, and [being] aware of what you're posting," which are both ongoing efforts. However, she later mentioned making the more singular choice to maintain private social media accounts because "it's more secure."

Understanding OPS as being agentive choices as opposed to hard and fast rules held important implications for the youth, particularly surrounding the idea of calculated risks. Youth across grade bands in this study reported knowing about poor OPS choices, including making weak passwords, talking to strangers, illegally streaming content, and visiting questionable websites. However, the youth—in particular the MS and HS youth—were still making these choices anyways after deciding that either the consequences were low, the reward was worth the risk, or both. For example, participant Y39HS admitted knowing that "pirating NBA basketball streams that go through lots of different ads and [involve] clicking off and stuff" was a "very, very bad" choice, but that he consistently chose to do it anyways because "it's the only way I can watch the games." Such descriptions of calculated risks highlighted a particular sort of self-aware confidence shared by most of the youth: in the pre-interview questionnaire, only about one-third of youth participants (30%) said they knew "a lot" about online privacy, with that number dropping to 10% (4 participants) saying the same about online security (see Appendix A for youth self-reported knowledge levels). However, only 3 participants (7.5%) admitted that they do not believe they use their devices securely, while 79.5% (29 participants) stated that they always use their devices securely. This apparent contradiction was summarized beautifully by participant Y27HS. When asked why she chose "a Moderate amount" for the questionnaire questions asking how much participants felt they knew about OPS, she replied: "I feel like I know enough. I might not know a lot, but I think I know enough of how to keep myself safe online."

4.2. RQ2: How do parents understand the role of OPS in youth's lives?

Regardless of how parents viewed OPS in their own lives—which was varied—they unanimously agreed that these concepts were important for their children. For example, P11HS viewed her own online privacy as "a trade-off" in which "the more they [i.e., Google] know about me, the more relevant content I feel I'm going to get." However, when it came to her child, she noted that "especially for a kid… he has to be extra careful." Her sentiments were echoed by all 40 parents, who worried specifically about the consequences of their children's actions. These concerns led to an emphasis on talking about the consequences of youth's poor behavior as opposed to focusing on the benefits of good behavior.

The parents' understanding of consequences were shared evenly across both online privacy and online security, with the most frequently cited consequences being hacking, future social or professional repercussions, data loss or misuse, kidnapping or stalking, theft, seeing inappropriate content, and mental health repercussions. Parents of ES children were more likely to mention the consequence of their child seeing inappropriate content, while parents of MS children were more likely to worry about their child experiencing mental health repercussions from online social interactions.

While parents, themselves, were quite worried about the consequences of poor OPS choices, they generally did not feel that youth were similarly concerned. Parents across grade bands stated a belief that OPS *should* or *would* matter to their children at some point, but that right now "it's just not something that they're thinking about" (P27HS) or are "as interested in" (P02MS). Parents of ES and some MS children felt that youth in these grades were too young to "necessarily think about the ramifications" (P33MS), or did not have enough high-stakes accounts or developmental knowledge yet for privacy and security to truly matter. For example, P16ES shared that youth her child's age make choices "based off of their desires and things they want [without connecting] it to 'this could affect your real life.'" These parents were more likely than their HS counterparts to say that privacy and security mattered some now, but that "as they get older...they'll start to get it (P10ES). Parents of upper MS and HS youth described youth in these grades as being more impulsive and explained that "at their age, they just want to be accepted" (P21MS) and "don't think about the consequences down the road" (P37HS). This impulsiveness, parents explained, was the root of OPS mistakes.

Interestingly, parents' beliefs about "kids that age" were only sometimes reflected in their opinions of their own children, resulting in what we have dubbed the "good kid syndrome." Parents experiencing "good kid syndrome" were those who gave conflicting responses about what "youth" do versus what they believe their own child does, believing that their child was more secure than most other children. Examples of parents with "good kid syndrome" included P13MS who stated that "I'm sure there are all kinds of kids who give their name to people they don't know, maybe other whether large or small pieces of information that could personally identify them," but when asked about whether his child has done the same stated "I'm sure she has, knowingly or unknowingly, but I think hers are probably, in my view, I think they're probably average to below average versus other kids" (P13MS). Similarly, P27HS suggested that many teenagers "[connect] with people on social media that they don't know personally" but that her child was "one of the good ones…a level-headed kid."

4.3. RQ3: How do parents attempt to influence youth's privacy and security understandings?

4.3.1 How parents monitor their children's online activities

Parents used a variety of methods to physically monitor their children's device use in an attempt to ensure that their children were private and secure online. The monitoring methods that were specifically mentioned included restricting access to devices (i.e., at night or as punishment), limiting screen time, controlling in-device purchasing and browsing using

parental monitoring applications, blocking websites or applications, observing device use, requiring devices to be used in a shared living space, and physically checking devices (i.e., looking at social media or browsing histories).

The parents who monitored their children ranged from passively monitoring using one method on an infrequent basis, to very actively monitoring, like P33MS who limited technology access, used parental controls, and observed use. Across all monitoring types, the amount and intensity of monitoring decreased as youth got older. This was found both within grade bands—HS parents reported decreasing monitoring behavior over time—as well as across all dyads, with ES parents reporting doing more monitoring than HS parents, and MS parents falling in between.

Parents of ES and MS youth were the most likely to rely on parental controls, but had complaints that parental controls were not nuanced enough, particularly for pre-teen youth who "kind of [fall] through the cracks…there's no in-between" (P18ES). Overall, despite the proliferation of device monitoring of all kinds, when we compared parents' monitoring choices with their children's understandings of OPS, we found no significant patterns between amount or type of monitoring and level of youth understanding.

4.3.2 How parents talk to their children

In addition to monitoring, many parents reported having conversations with their children about OPS. A majority of the conversations that parents described having were about the consequences of poor choices, and were reactionary in nature. For example, P30HS recalled having a conversation about talking with strangers online and security settings *after* her daughter and a friend "were playing Roblox and a weirdo, an adult male, decided to chat with them." Similarly, P02MS admitted talking more about online privacy than security with her daughter because "that's where I've seen the issue, honestly."

Parents who chose *not* to have conversations about OPS with their children felt that the knowledge was coming from elsewhere, like P22HS said she did not have OPS conversations with her son because "I think he's been given lessons about it in school." Technology was also cited as a reason to not need to have explicit conversations, like when P28HS explained that she "rel[ies] on a lot of websites that require a capital and a lowercase and a number [for passwords], so that kind of takes care of it," and P35HS reasoned that "if I have this [security] suite and I keep it up to date, that should generally protect him…I've not had a conversation with him."

Parents' decisions to have conversations about OPS dependent on their child's age. Parents of ES and MS youth most frequently reported either tailoring their conversations to their child's perceived technological understandings or holding off on the conversation altogether until their children are older. P10ES explained that she currently only has a "small amount" of conversations with her daughter "due to age and because she only has a tablet…but as [she] gets older and [she gets] more independent, of course, you need to have those conversations." Similarly, while P17ES made and stored all of her son's passwords for him at the time of interview, she noted that "sometime soon [he's] going to have to pick his own password for something…[and] then he'll probably listen and we'll discuss it."

Conversely, the parents of HS youth believed that their children either already know about online privacy or were old enough for the conversations to no longer be necessary. P11HS was one such mom, who described her son as "a little man," and noted that "we've had all those conversations, but it's been years. I honestly don't know what he knows at this point…because honestly we haven't had those conversations probably in three or four years." Interestingly this parent, as well as several of her peers who reported not talking to their children about these topics, overwhelmingly stated that they (the parents) were most responsible for their children's knowledge, while also admitting that they do not regularly (if ever) talk to their children about OPS.

Finally, the results of this study reveal that parents want to know more about OPS but are unsure how to do so. Of the parents who described not talking to their children about OPS at all, all but one self-reported knowing "little" about either online privacy, security, or both. One of these parents reflected that "this research has reminded me how little I know about OPS, and since my kids are young, it's my job to teach them" (P12ES). P31MS noted: "I hope that as a parent I can stay on top of all the changing online interactions…[but] I feel a bit overwhelmed at times regarding this topic." These comments, combined with the large number of parents who said they "want to learn more about this topic and how I can make better safety choices for me and my children" (P35HS) suggest that parents' perceived levels of knowledge may impact the amount and kind of conversations they choose to have with their children about OPS. This possibility is especially interesting considering only 4 parents in this study (10%) reported feeling like they know "a lot" about OPS.

4.3.3 The Influence of Parents on Children

Overall, the youth in this study with more nuanced descriptions and understandings of OPS had parents who reported having conversations with them instead of or in addition to the monitoring of device use. This finding held true regardless of the parents' self-reported levels of OPS knowledge (see Appendix A), of how confident they were in having the conversations, or of the strength of parents' own stated understandings. For example, when talking about having online security discussions with her son, P20MS explained that "we just basically talk to him about the fact that certain websites are inappropriate or even could give him a virus." She also

noted that the conversations involved explanations and were "usually not just like 'oh we don't want you to,' we usually give him a reason why we don't want him to be on that and why that behavior is inappropriate" (P20MS). These conversations were directly reflected in the ways Y20MS talked about online security: he mentioned keeping information secure using VPNs and firewalls, avoiding pop-ups and dangerous websites that can cause viruses, noted that he runs security scans on his computer, and discussed the role of personal choice in a person's level of online security. He also identified the consequences of risky online behavior as including "leading you to something that's really inappropriate" and "giving your computer malware or a virus" (Y20MS).

By contrast, the youth—especially ES and MS aged youth—whose parents relied on physical monitoring in lieu of conversations could generally provide descriptions of OPS, but struggled to explain *why* OPS is important and to provide examples. For example, P26ES noted that she "monitor[s] [her son's] phone to the fifth power" and, when asked if they have conversations about privacy and security, replied yes. After replying yes, however, she proceeded to provide an example of hearing foul language during a video game at which time she "took his headphones and said, 'you can't play, just turn it off.'" Consequently, her fifth-grade son Y26ES described online privacy as "not to be bothered" online, and said he did not know what online security was.

With ES aged youth in particular, the parents who reported having conversations with their child had youth with more nuanced understandings. Conversely, elementary aged youth like Y26ES whose parents chose not to discuss OPS with them demonstrated lower levels of understanding and less nuanced descriptions. With HS participants, this gap disappeared: HS parents almost unanimously reported not having recent online privacy or security conversations with their children, but most HS youth still provided detailed descriptions and nuanced examples of both terms.

5. Discussion

Our study sought to learn more about youth's OPS knowledge, as well as how parents understand and attempt to influence that knowledge. It was unique in its design and purpose in two ways. First, we studied parent/youth dyads (as opposed to one population or the other) with a broader age range of participants (10 standard United States school grades–3rd to 12th), which allowed us to examine findings both within and across dyads and grade bands to look for interrelationships not able to be studied using other designs. Second, we were also curious about youth knowledge in general versus in response to specific learning interventions or experiences.

In terms of youth OPS knowledge, when asked to describe and give examples of OPS, the 40 youth in our study, regardless of age, were able to describe both terms, and were able to name examples and online choices that exemplified good and bad OPS behavior. Their descriptions and examples supported several preexisting findings about youth's OPS understandings, particularly that youth do know about, care about, and value these ideas [67]; and that there are often gaps between what youth say they know and the actions they take ([15][44]). Our study also supported existing findings about parents' understanding of their children's knowledge and attempts to influence that knowledge, namely that parents frequently choose device monitoring and physical or technical controls over conversations ([27][64]); often hold misguided understandings about youth's knowledge, including the idea that younger children are too young to understand or exercise protective practices[40]; and are concerned about their own knowledge not being strong enough to best support their children [15]. Our study's most compelling findings arose when examining youth and parent knowledge both within and across dyads and grade bands. Our study's greatest contribution to the ongoing investigation of youth OPS knowledge is our examination of the relationships between parent knowledge, parent OPS monitoring and education, and youth knowledge .

5.1. Parental Influence on Youth Understandings

What the 40 parents in this study understood about their children's OPS knowledge can be summarized into three broad categories: those who believed their children were too young to fully understand or care, those who believed their children were "good kids" who wouldn't get into trouble, and those who felt like their children already knew enough to make good choices. All three beliefs, however, generated similar parental influence responses: an emphasis on passive monitoring (i.e., parental controls and device monitoring), or conversations that mostly centered on consequences of poor online choices. We also found, however, that parental conversations—either alone or in conjunction with monitoring—may be more effective at establishing stronger youth understandings than device monitoring alone.

These conflicting factors—along with the fact that the parents of younger youth frequently mentioned that conversations with their children would happen "later," while parents of older youth noted that such conversations are no longer necessary—collectively raise the question of when the magic time frame for conversations with children about OPS is, and if these conversations ever wind up consistently happening at all. On one hand, younger youth whose parents did more passive monitoring than conversational engagement had less nuanced privacy and security understandings. On the other hand, high school youth had more complete and nuanced understandings regardless of the methods of parental influence. This suggests that at some point, children begin gaining OPS knowledge from sources outside their parents that help round

out their understandings. However, up until that point, parental influence has the potential to make a meaningful difference in youth OPS knowledge and behavior, and the type of parental influence matters.

Further, youth in this study understood OPS as agentive and user-influenced, suggesting that conversations with youth about decision-making surrounding the use and sharing of information and data online may be more important than more prescriptive approaches to building understanding like defining rules or pre-setting controls. If youth understand OPS as elements of their technical selves that require risk calculation and choice, having conversations about how to weigh such decisions and make good choices—as well as the potential consequences of choosing to engage in less private and less secure choices—is likely more helpful than monitoring. As youth grow and their online activities diversify, they will be increasingly faced with choices concerning their OPS and need to be armed with the knowledge and skills to make these choices, and parents simply cannot always be watching. The youth and parents in our study indicate that parental conversations with youth either in addition to other forms of monitoring or as the sole form of monitoring, alone, is likely a better approach. Further, contrary to parental belief, there is no such thing as "too early" for these conversations because, as these 40 youth indicate, youth of *all* ages understand the importance of OPS and are prepared to think about how to protect themselves online.

5.2. Implications

5.2.1 Implications for Parents

The primary takeaway from this study for parents is straightforward: talk about OPS choices with children, and begin doing so in the elementary years as soon as youth are given access to devices. Our study suggests that parents do not have to be experts—or even be incredibly confident in their own OPS knowledge—for these conversations to be successful. Rather, especially given the ever-evolving nature of these topics [18], parents can co-construct and continue to learn alongside their youth via conversations about OPS choices and behaviors versus feeling like they need to be OPS experts to be helpful. This idea of co-constructing knowledge could help overcome the gaps in knowledge that both youth *and* parents have when it comes to OPS ([8][66]), as well as prepare youth to be informed decision-makers when making OPS choices they feel they are responsible for.

5.2.2 Implications for Technology Providers

Like other literature examining parent and youth understandings of privacy, our study supported that both youth and parents think about online security and especially online privacy more at the interpersonal levels and less at the commercial and institutional levels [33]. It also revealed through a social learning lens that by middle school, most youth may be getting as much or more of their information about OPS from outside the home, including from technology tools, devices, applications, and services. This means that technology providers have the opportunity and possibly even the responsibility to support youth knowledge, especially when it comes to understandings like how data is collected, stored, and tracked. These providers might consider making more proactive, outcome-based tools to support parents instead of monitoring-based ones, or creating more educational tools to teach young users about OPS choices and choice-outcomes. Similarly, providers might consider creating tools for passive-monitoring parents to help them supplement their current strategies with conversational approaches.

5.2.3 Implications for Future Research

Extant research and literature tend to either conflate online privacy and security, or to specifically investigate one of the terms in isolation from the other. Our study—which investigated both terms separately from each other—reveals that both parents and youth of all ages do understand these terms as interrelated but distinct, and that youth have more knowledge and exposure to online privacy than online security. Future user-centered research should further explore youth's interconnected understandings to explore how youth use their OPS knowledge in conjunction to stay secure and private online instead of as separate or singular entities. Further, our study showed that especially older youth approach OPS from an agentive perspective, and intentionally make choices to engage or not engage in private and secure behavior. More research investigating the nature of these choices and how youth make them could go a long way in continuing to support our understanding of youth habits.

Finally, our study preliminarily reveals that the when, how, and what of parent conversations about OPS has the power to influence youth understandings, especially with youth in elementary and middle school. Further qualitative explorations into the kinds of conversations that parents have could help build a better understanding of exemplary characteristics of such conversations. More dyadic studies are a recommended approach to this work because of their unique ability to examine both parent and youth actions, perspectives, and resulting knowledge. Importantly, work is needed to understand how parents and youth feel about the extent and type of influence each have on understandings of OSP.

6. Positionality and Limitations

The positionality of the authors in this study is important to note. The first author is an educator and has worked with youth for over a decade and the second author is a parent, and these positionalities and related experiences explicitly surfaced throughout data analysis discussions. The influences of these positionalities were mitigated by a rigorous data analysis process, as well as full-team data analysis discussions during which individual assumptions surfaced and were

identified and discussed by the research team, including the third author who was neither a parent nor a teacher.

Additionally, this study had several limitations. First, because of the COVID-19 pandemic, interviews happened via video platform. This meant that, although we requested that individual participants complete the interview privately, there was a nonzero chance of youth/parent participants overhearing and influencing each other's responses or responding with the possibility of being overheard by others. Further, common limitations for qualitative data in general applied to our study, including the possibility of biases in participants' self-reports of behavior (e.g. optimism bias [48]), and potential order effects by asking about online privacy first and security second [39].

Finally, our study was limited by its cross-sectional design focusing only on parents and youth at one point in time. This study was not longitudinal, meaning we could compare dyads within and across age groups, but could not examine the progression of parental influence and youth knowledge of the same dyads over time. Further, our theoretical approach requires an understanding that youth knowledge is impacted by a variety of factors including things like school and peers, but we scoped the study specifically to the influence of parents.

Each of these design limitations offers important potential directions for future research, such as focusing on longitudinal data and/or more holistic approaches to understanding how a variety of factors are influencing youth at different ages.

7. Conclusion

In conclusion, this study showed that the 40 3rd-12th grade youth we interviewed had at least basic—and at times nuanced and interconnected—OPS knowledge. Further, they viewed these topics as important and were aware that participation in the online world includes frequent opportunities to make privacy- and security-related choices. For these youth, particularly the younger youth in elementary and middle school, parents were influential contributors to this knowledge and these choices, and had the power and influence to help their children be more private and secure online. For these 40 parents, the most effective strategy for influencing their children's knowledge and understandings was through conversations and learning alongside their children, even when they thought their children might not be interested or old enough to fully "get it."

Continued learning about what young users know about OPS is an important step in the ongoing process of discovering how, when, and where to teach them this knowledge. The younger that youth can learn to flexibly practice strong OPS practices, the better prepared they will be to keep themselves secure online. Further, because we know that parents play an active and important role in this learning [4], the more we can help prepare parents to have constructive conversations about OPS with their children the better.

References

[1] Fenio Annasingh and Thomas Veli. An investigation into risks awareness and e-safety needs of children on the internet. *Interactive Technology and Smart Education*, vol. 13, no. 2, pp. 147-165, 2016.

[2] Lioness Ayres. Semi-structured interview. *The SAGE encyclopedia of qualitative research methods*, vol. 1, pp. 810-811, 2008.

[3] Karla Badillo-Urquiola, Yaxing Yao, Oshrat Ayalon, Bart Knijnenurg, Xinru Page, Eran Toch, Yang Wang, and Pamela J. Wisniewski. Privacy in Context: Critically Engaging with theory to guide privacy research and design. In *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing,* 2018, October, pp. 425-431.

[4] Albert Bandura. *Social Learning Theory,* New York: General Learning Press, 1977.

[5] Stacy Black, Rezvan Joshaghani, Dhanush Kumar Ratakonda, Hoda Mehrpouyan, and Jerry Alan Fails. Anon what what? Children's Understanding of the Language of Privacy. In *Proceedings of the 18th ACM International Conference on Interaction Design and Children*, pp. 439-445, 2019.

[6] Hennie Boeije. A purposeful approach to the constant comparative method in the analysis of qualitative interviews. *Quality and quantity,* vol. *36,* no. 4, pp. 391-409, 2002.

[7] Laura Brandimarte, Alessandro Acquisti, and George Loewenstein. Misplaced confidences: Privacy and the control paradox. *Social Psychological and Personality Science*, vol. 4, no. 3, pp. 340-347, 2013.

[8] Eva Irene Brooks and Anders Kalsgaard Moeller. Children's perceptions and concerns of online privacy. *Extended Abstracts of the Annual Symposium on Computer-Human Interaction in Play Companion Extended Abstracts*, pp. 357-362, 2019.

[9] Sangmi Chai, Sharmistha Bagchi-Sen, Claudia Morrell, H. Raghav Rao, and Shambhu J. Upadhyaya. Internet and online information privacy: An exploratory study of preteens and early teens. *IEEE Transactions on Professional Communication*, vo. 52, no. 2, pp. 167-182, 2009.

[10] Hui-Lien Chou, Yih-Lan Liu, and Chien Chou. Privacy behavior profiles of underage Facebook users. *Computers & Education,* vol.128, pp. 473-485, 2019.

[11] Merijke Coenraad, Anthony Pellicone, Diane Jass Ketelhut, Michel Cukier, Jan Plane, and David

Weintrop. Experiencing cybersecurity one game at a time: A systematic review of cybersecurity digital games. *Simulation & Gaming*, vol. 51, no. 5, pp. 586-611, 2020.

[12] Katie Davis, and Carrie James. Tweens' conceptions of privacy online: Implications for educators. *Learning, Media and Technology*, vol. 38, no. 1, pp. 4-25, 2013.

[13] John Dempsey, Gavin Sim, and Brendan Cassidy. Designing for GDPR-investigating children's understanding of privacy: A survey approach. *32nd Human Computer Interaction Conference,* Belfast, Ireland, 2018.

[14] Laura M. Desimone, and Kerstin Carlson Le Floch. Are we asking the right questions? Using cognitive interviews to improve surveys in education research. *Educational evaluation and policy analysis*, vol. 26, no. 1, pp. 1-22, 2004.

[15] Laurien Desimpelaere, Liselot Hudders, and Dieneke Van de Sompel. Knowledge as a strategy for privacy protection: How a privacy literacy training affects children's online disclosure behavior. *Computers in human behavior*, vol. 110, no. 106382, 2020.

[16] Sonya Corbin Dwyer and Jennifer L. Buckle. The space between: On being an insider-outsider in qualitative research. *International journal of qualitative methods* vol. 8, no. 1, pp. 54-63, 2009.

[17] Yang Feng and Wenjing Xie. Teens' concern for privacy when using social networking sites: An analysis of socialization agents and relationships with privacy-protecting behaviors. *Computers in Human Behavior,* vol. 33, pp. 153-162, 2014

[18] David Finkelhor, Lisa Jones, and Kimberly Mitchell. Teaching privacy: A flawed strategy for children's online safety. *Child Abuse & Neglect* vol. 117, no. 105064, 2021.

[19] Alexa K. Fox and Mariea Grubbs Hoy. Smart devices, smart decisions? Implications of parents' sharenting for children's online privacy: An investigation of mothers. *Journal of Public Policy & Marketing*, vol. 38, no. 4, pp. 414-432, 2019.

[20] Steven Furnell and Emily Collins.S. Cyber security: what are we talking about? *Computer Fraud & Security*, vol. 2021, no. 7, pp. 6–11, 2021, doi: https://doi.org/10.1016/S1361-3723(21)00073-7.

[21] Steven Furnell, Rawan Esmael, Weining Yang, and Ninghui Li. Enhancing security behaviour by supporting the user. *Computers & Security,* vol. 75, pp. 1-9, 2018.

[22] Simon L. Jones, Emily IM Collins, Ana Levordashka, Kate Muir, and Adam Joinson. What is 'Cyber Security'? Differential Language of Cyber Security Across the Lifespan. in *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–6. Doi: 10.1145/3290607.3312786

[23] Flavius Kehr, Tobias Kowatsch, Daniel Wentzel, and Elgar Fleisch. Blissfully ignorant: the effects of general privacy concerns, general institutional trust, and affect in the privacy calculus. *Information Systems Journal*, vol. 25, no. 6, pp. 607-635, 2015.

[24] Beeban Kidron, Anghrarad Rudkin, Miranda Wolpert, Joanna R. Adler, Andrew K. Przybylski, Elvira Perez Vallejos, Henrietta Bowden-Jones, Joshua J. Chauvin, Kathryn L. Mills, Marina Jirotka, and Julian Childs. *Digital Childhood: Addressing Childhood Development Milestones in the Digital Environment,* Technical Report, 5Rights, 2017.

[25] Abdullah Konak. Experiential learning builds cybersecurity self-efficacy in K-12 students. *Journal of Cybersecurity Education, Research and Practice,* no. 2018(1), p. 6, 2018.

[26] Priya Kumar, Shalmali Milind Naik, Utkarsha Ramesh Devkar, Marshini Chetty, Tamara L. Clegg, and Jessica Vitak. 'No Telling Passcodes Out Because They're Private' Understanding Children's Mental Models of Privacy and Security Online. *Proceedings of the ACM on Human-Computer Interaction* 1, no. CSCW, pp. 1-21, 2017.

[27] Priya Kumar, Jessica Vitak, Marshini Chetty, Tamara L. Clegg, Jonathan Yang, Brenna McNally, and Elizabeth Bonsignore. Co-designing online privacy-related games and stories with children. In *Proceedings of the 17th ACM conference on interaction design and children*, pp. 67-79, 2018.

[28] J. Richard Landis, and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, pp. 159-174, 1977.

[29] Sonia Livingstone. Children's privacy online: experimenting with boundaries within and beyond the family. In R. Kraut, M. Brynin, and S. Kiesler (eds.) *Computers, Phones, and the Internet: Domesticating Information Technology,* Human technology interaction series, Oxford University Press, New York, pp. 145-167, 2006.

[30] Sonia Livingstone, Giovanna Mascheroni, Michael Dreier, Stephane Chaudron, and Kaat Lagae. *How parents of young children manage digital devices at home: The role of income, education and parental style*, London: EU Kids Online, LSE, 2015.

[31] Sonia Livingstone, Giovanna Mascheroni, and Elisabeth Staksrud. European research on children's internet use:

Assessing the past and anticipating the future," *New media & society*, vol. 20, no. 3, pp. 1103-1122, 2018.

[32] Sonia Livingstone and Kjartan Olafsson. When do parents think their child is ready to use the internet independently? *Parenting for a digital future: Survey Report 2, Department of Media and Communications*, the London School of Economics and Political Science, London, UK, 2018.

[33] Sonia Livingstone, Mariya Stoilova, and Rishita Nandagiri. *Children's data and privacy online: growing up in a digital age: an evidence review*, London School of Economics and Political Science, Department of Media and Communications, London, UK, 2019.

[34] Paweena Manotipya and Kambiz Ghazinour. Children's Online Privacy from Parents' Perspective. *Procedia Computer Science,* vol. 177, pp. 178-185, 2020.

[35] Craig McDonald-Brown, Kumar Laxman, and John Hope. An exploration of the contexts, challenges and competencies of pre-teenage children on the internet. *International Journal of Technology Enhanced Learning*, vol. 8, no. 1, pp. 1-25, 2016.

[36] Mary L. McHugh. Interrater reliability: the kappa statistic. *Biochemia medica,* vol. 22, no. 3, pp. 276-282, 2012.

[37] Ingrida Milkaite, Ralf De Wolf, Eva Lievens, Tom De Leyn, and Marijn Martens. Children's reflections on privacy and the protection of their personal data: A child-centric approach to data protection information formats. *Children and Youth Services Review,* vol. 129, 2021.

[38] Tehila Minkus, Kelvin Liu, and Keith W. Ross. Children seen but not heard: When parents compromise children's online privacy. In *Proceedings of the 24th International Conference on World Wide Web*, pp. 776-786, 2015.

[39] David W. Moore. Measuring new types of question-order effects: Additive and subtractive. *The Public Opinion Quarterly*, vol. 66, no. 1, pp.80-91, 2002.

[40] James Nicholson, Julia Terry, Helen Beckett, and Pardeep Kumar. Understanding Young People's Experiences of Cybersecurity. *European Symposium on Usable Security,* pp. 200-210, 2021.

[41] Joshua C. Nwokeji, Richard Matovu, and Bharat Rawal. The use of Gamification to Teach Cybersecurity Awareness in information systems. In *Proceedings of the 2020 AIS SIGED International Conference on Information Systems Education and Research,* no. 29, 2020.

[42] Sean Oesch, Ruba Abu-Salma, Oumar Diallo, Juliane Krämer, James Simmons, Justin Wu, and Scott Ruoti. User Perceptions of Security and Privacy for Group Chat: A Survey of Users in the US and UK. *Annual Computer Security Applications Conference*. Association for Computing Machinery, Austin, USA, pp. 234–248, 2020. Doi: 10.1145/3427228.3427275.

[43] Gwenn Schurgin O'Keeffe, and Kathleen Clarke-Pearson. The impact of social media on children, adolescents, and families. *Pediatrics*, vol. 127, no. 4, pp. 800-804, 2011.

[44] Luci Pangrazio and Neil Selwyn. 'My Data, My Bad...' Young People's Personal Data Understandings and (Counter) Practices. In *Proceedings of the 8th International Conference on Social Media & Society*, pp. 1-5, 2017.

[45] Jean Piaget. Intellectual evolution from adolescence to adulthood. *Human development*, vol. 15, no. 1, 1-12, 1972.

[46] Farzana Quayyum, Daniela S. Cruzes, and Letizia Jaccheri. Cybersecurity awareness for children: A systematic literature review. *International Journal of Child-Computer Interaction*, vol. 30, p. 100343, 2021.

[47] Johnny Saldaña. *The coding manual for qualitative researchers,* 3rd ed., Sage, 2016.

[48] Tali Sharot. The optimism bias. *Current biology*, vol. 21, 23, pp.941-945, 2011.

[49] Wonsun Shin, and Hyunjin Kang. Adolescents' privacy concerns and information disclosure online: The role of parents and the Internet. *Computers in Human Behavior,* vol. 54, pp. 114-123, 2016.

[50] Kingkarn Sookhanaphibarn and Worawat Choensawat. Educational Games for Cybersecurity Awareness. *2020 IEEE 9th Global Conference on Consumer Electronics (GCCE)*, pp. 424-428, IEEE, 2020.

[51] M. Stoilova, S. Livingstone, S., and R. Nandagiri, *Children's data and privacy online: Growing up in a digital age, Research findings,* London: London School of Economics and Political Science, 2019.

[52] Mariya Stoilova, Sonia Livingstone, and Rishita Nandagiri. Digital by default: Children's capacity to understand and manage online data and privacy. *Media and Communication*, vol. 8, no. 4, pp. 197-207, 2020.

[53] Katrien Symons, Koen Ponnet, Michel Walrave, and Wannes Heirman. A qualitative study into parental mediation of adolescents' internet use. *Computers in Human Behavior,* vol. 73, pp. 423-432, 2017.

[54] Monika Sziron and Elisabeth Hildt. Digital media, the right to an open future, and children 0–5. *Frontiers in Psychology*, 2137, 2018.

[55] Edwin Van Teijlingen and Vanora Hundley. The importance of pilot studies. *Social research update*, vol. 35, no. 4, pp. 49-59, 2010.

[56] Mary Theofanos, Yee-Yin Choong, and Olivia Murphy. 'Passwords Keep Me Safe'–Understanding What Children Think about Passwords. *30th USENIX Security Symposium (USENIX Security 21)*, pp. 19-35. 2021.

[57] Sreenivas Sremath Tirumala, Abdolhossein Sarrafzadeh, and Paul Pang. A Survey on Internet Usage and Cybersecurity Awareness in Students. *14th Annual Conference on Privacy, Security and Trust (PST),* 2016.

[58] U.S. Department of Education, National Center for Education Statistics. Children's Internet Access at Home. [Online], *The Condition of Education 2021* (NCES 2021-144), 2021, Available: https://nces.ed.gov/programs/coe/indicator/cch

[59] Martin Valcke, Tammy Schellens, Hilde Van Keer, and Marjan Gerarts. Primary school children's safe and unsafe use of the Internet at home and at school: An exploratory study. *Computers in human behavior*, vol. 23, no. 6, pp. 2838-2850, 2007.

[60] José Van Dijck. *The culture of connectivity: A critical history of social media*. Oxford University Press, 2013.

[61] Jacqueline M. Williams-Reade, Daniel Tapanes, Brian J. Distelberg, and Susanne Montgomery. Pediatric Chronic Illness Management: A Qualitative dyadic analysis of adolescent patient and parent illness narratives. *Journal of marital and family therapy*, vol. *46*, no. 1, pp. 135-148, 2020.

[62] Christine Ee Ling Yap and Jung-Joo Lee. 'Phone apps know a lot about you!' educating early adolescents about informational privacy through a phygital interactive book. *Proceedings of the Interaction Design and Children Conference*, pp. 49-62, 2020.

[63] Ramazan Yilmaz, F. Gizem Karaoglan Yilmaz, H. Tugba Özturk, and Tugra Karademir. Examining secondary school students' safe computer and internet usage awareness: an example from Bartın province. *Pegem Journal of Education and Instruction,* vol. 7, no. 1, pp. 83-114, 2017.

[64] Leah Zhang-Kennedy, Christine Mekhail, Yomna Abdelaziz, and Sonia Chiasson. From nosy little brothers to stranger-danger: Children and parents' perception of mobile threats. In *Proceedings of the The 15th International Conference on Interaction Design and Children*, pp. 388-399, 2016.

[65] Leah Zhang-Kennedy, Yomna Abdelaziz, and Sonia Chiasson. Cyberheroes: The design and evaluation of an interactive ebook to educate children about online privacy. *International Journal of Child-Computer Interaction,* vol. 13, pp.10-18, 2017.

[66] Jun Zhao. Are Children Well-Supported by Their Parents Concerning Online Privacy Risks, and Who Supports the Parents? *arXiv preprint arXiv:1809.10944*, 2018.

[67] Jun Zhao, Ge Wang, Carys Dally, Petr Slovak, Julian Edbrooke-Childs, Max Van Kleek, and Nigel Shadbolt. I make up a silly name' Understanding Children's Perception of Privacy Risks Online. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1-13, 2019.

[68] Zahidah Zulkifli, Nurul Nuha Abdul Molok, Nurul Hayani Abd Rahim, and Shuhaili Talib. Cyber Security Awareness Among Secondary School Students In Malaysia. *Journal of Information Systems and Digital Technologies,* vol. 2, no. 2, pp. 28-41, 2020.

Appendix A. Demographics and self-reported OPS knowledge responses

Dyad ID	Youth (Grade)	Youth Self-Rated Knowledge		Youth Self-Reported whether always using electronic devices securely	Parent (Age)	Parent Self-Rated Knowledge		Parent Perception on whether the child always uses electronic devices securely
		Privacy	Security			Privacy	Security	
D01	Boy (8th)	A lot	A lot	Yes	Mom (48)	Moderate	Moderate	Yes
D02	Girl (6th)	Moderate	Moderate	Yes	Mom (48)	Moderate	Moderate	No
D03	Boy (6th)	Moderate	Moderate	Yes	Mom (38)	Moderate	Moderate	Yes
D04	Girl (9th)	Little	Little	Not sure	Mom (44)	Moderate	Moderate	No
D05	Boy (12th)	Little	Little	Yes	Mom (37)	Moderate	Moderate	Yes
D06	Boy (3rd)	A lot	A lot	Yes	Mom (34)	Moderate	Moderate	Yes
D07	Boy (4th)	Moderate	Little	Yes	Mom (39)	Little	Little	Yes
D08	Boy (3rd)	Little	Little	Not sure	Mom (37)	Moderate	Moderate	Not sure
D09	Girl (4th)	Moderate	Little	Yes	Dad (35)	A lot	A lot	No
D10	Girl (3rd)	Little	Little	Yes	Mom (31)	Moderate	Moderate	Yes
D11	Boy (10th)	Moderate	Moderate	No	Mom (52)	Moderate	Moderate	Not sure
D12	Girl (5th)	Moderate	Moderate	Yes	Mom (50)	Little	Little	No
D13	Girl (7th)	Moderate	Moderate	Yes	Dad (46)	Moderate	Moderate	No
D14	Boy (4th)	A lot	Moderate	Yes	Mom (52)	Moderate	Moderate	Yes
D15	Girl (5th)	Moderate	A lot	Yes	Mom (34)	Moderate	Moderate	Not sure
D16	Girl (4th)	A lot	Moderate	Yes	Mom (39)	Moderate	Little	No
D17	Boy (3rd)	-	-	Not sure	Dad (41)	Moderate	Moderate	No
D18	Girl (5th)	A lot	Moderate	Yes	Mom (52)	Moderate	Moderate	No
D19	Boy (7th)	A lot	Moderate	Yes	Dad (42)	Moderate	Moderate	Not sure
D20	Boy (8th)	A lot	Moderate	Yes	Mom (39)	Moderate	Moderate	No
D21	Girl (7th)	Little	Little	Yes	Mom (36)	A lot	A lot	No
D22	Boy (11th)	Moderate	Moderate	Yes	Mom (48)	Little	Little	Not sure
D23	Boy (9th)	A lot	Moderate	Yes	Mom (47)	Moderate	Moderate	No
D24	Boy (11th)	Moderate	-	Yes	Mom (36)	Moderate	Moderate	No
D25	Girl (8th)	Moderate	Moderate	Yes	Mom (33)	Moderate	Little	No
D26	Boy (5th)	Moderate	Little	No	Mom (42)	A lot	A lot	Yes
D27	Girl (11th)	Moderate	Moderate	Yes	Mom (39)	Moderate	Moderate	Not sure
D28	Girl (12th)	Moderate	Moderate	Not sure	Mom (41)	Little	Little	Not sure
D29	Boy (6th)	Little	Moderate	Not sure	Mom (34)	Little	Little	Not sure
D30	Girl (9th)	A lot	A lot	Yes	Mom (51)	A lot	A lot	No
D31	Girl (8th)	A lot	Little	Yes	Mom (45)	Moderate	Moderate	Yes
D32	Girl (6th)	A lot	Moderate	Yes	Mom (44)	Moderate	Moderate	Not sure
D33	Boy (7th)	Moderate	Little	Yes	Mom (50)	Moderate	Moderate	Yes
D34	Boy (10th)	Moderate	Little	Yes	Dad (51)	Moderate	Moderate	Not sure
D35	Boy (9th)	Moderate	Moderate	Not sure	Mom (42)	Moderate	Moderate	No
D36	Girl (12th)	Moderate	Moderate	Not sure	Mom (44)	Little	Little	Not sure
D37	Girl (10th)	A lot	Moderate	Yes	Mom (51)	Little	Little	No
D38	Girl (11th)	Moderate	Moderate	Not sure	Mom (48)	Little	Little	Yes
D39	Boy (12th)	Moderate	Little	No	Mom (39)	Little	Little	Not sure
D40	Boy (9th)	-	-	-	Mom (35)	Moderate	Little	Yes

Appendix B. Study Instruments

Pre-interview Questionnaire – Youth

1. Choose your gender: [radio buttons: *Boy/Girl* for ES; *Male/Female* for MS/HS]
2. How old are you? [entry field] (in years)
3. What is your grade? [drop-down list from 3rd to 12th]
4. Do you have a smartphone? [radio buttons: *Yes, my own/Yes, I share one with someone else/No*]
   [If the answer to Q4 is "No," then skip to Q5]
   4.1 How old were you when you first got a smartphone? [entry field] (in years)
   4.2 On average, how many hours a day do you spend on your smartphone? [entry field] (in hours)
5. Do you use a computer at home? (meaning a desktop, a laptop, or a tablet) [radio buttons: *Yes, my own/Yes, I share one with someone else/No*]
6. How would you define online privacy? [text area]
7. How much do you know about online privacy? [radio buttons: *A little/A middle amount/A lot* for ES; *Very little to nothing/A moderate amount/A lot* for MS/HS]
8. [ES] Who taught you about online privacy? [MS/HS] From whom did you learn about online privacy? [matrix with Yes/No options for each item below]
   Your parents or guardians; Brothers/Sisters for ES, *Siblings* for MS/HS]; *Other family members; Teachers/school; Friends; Yourself; Other*
9. How would you define online security? [text area]
10. How much do you know about online security? [radio buttons: *A little/A middle amount/A lot* for ES; *Very little to nothing/A moderate amount/A lot* for MS/HS]
11. [ES] Who taught you about online security? [MS/HS] From whom did you learn about online privacy? [matrix with Yes/No options for each item below]
    Your parents or guardians; Brothers/Sisters for ES, *Siblings* for MS/HS]; *Other family members; Teachers/school; Friends; Yourself; Other*
12. Do you think you always use your electronic device(s) securely? (meaning smartphone, desktop, laptop, tablet) [radio buttons: *Yes/No/I'm not sure*]
13. How would you define risky online behavior? [text area]
14. Have your parents/guardians spoken to you about risky online behaviors? [radio buttons: *Yes/No/I don't remember*]
15. Would you say you know more, the same, or less about technology than your parents/guardians? [radio buttons: *More/About the same/Less/I'm not sure*]

Semi-Structured Interview Scrip – Youth

1. Tell me about how you spend most of your time online.
2. I see you said you know [*response from questionnaire Q15*] about technology than your parents; can you explain this answer and why you said this?
3. (ES) Does anyone in your house watch or check in on what you do online? Does anyone control how much time you spend online? [If yes, who and how?] (MS/HS) Does anyone in your house monitor what you do online or how long you spend online? [If yes, who and how?]
4. I see you defined online privacy as [*response from questionnaire Q6*]. Do you think online privacy is important? [Why or why not?]
5. Give me an example or two of a good online privacy choice. [What about a bad privacy choice?][What do you think happens when someone makes a bad online privacy choice?]
6. I see you defined online security as [*response from questionnaire Q9*]. Do you think online security is important? [Why or why not?]
7. Give me an example or two of a good online security choice. [What about a bad security choice?][What do you think happens when someone makes a bad online security choice?]
8. I see you defined an online risk as [*response from questionnaire Q13*]. What are some examples of risky online behavior? Why do you think people take online risks? What happens to people who do [*repeat answers child just gave about online risks*]? Why do you think people make risky choices even if they know they're risky?
9. Can you remember making any risky choices online you can tell me about? Did you know they were risky at the time? What happened because of those risky choices?
10. What are the most important things you can do to stay private and secure online?
11. Who do you think is most responsible for keeping you private and secure online?

Pre-interview Questionnaire – Parents

1. What is your relationship to your child? [radio buttons: *Mom/Dad/Other relative or non-family guardian (describe)*]
2. What is your gender? [radio buttons: *Male/Female*]
3. What is your age? [entry field] (in years)
4. What is your highest level of education? [radio buttons: *Some high school/High school diploma/Some college/bachelor's degree/Master's degree/Doctoral degree/Other (specify)*]
5. What is your occupation? [text area]
6. How many children under 18 live in your household? [text area]
7. How many people in total live in your household? [text area]
8. In general, when does your household adopt new technologies? [radio buttons: *We try the latest technologies as soon as they come out/We follow technology trends/We let others work out the kinks first/We wait until our old technology dies/We wait until new technology becomes affordable for us*]
9. Do you own a smartphone? [radio buttons: *Yes/No*]
10. Does your child have their own or share a smartphone? [radio buttons: *Yes, their own/Yes, they share one/No*]
    [If the answer to Q10 is "No," then skip to Q11]
    10.1 At what age did you first give your child access to a smartphone? [entry field] (in years)
    10.2 On average, how many hours a day do you believe your child spends on a smartphone? [entry field] (in hours)
11. Does your child have access to computer(s) in your home? [radio buttons: *Yes, they have their own device/Yes, they share one with me or other family members/No*]
    [If the answer to Q11 is "No," then skip to Q12]
    11.1 At what age did you first give your child access to computers at home? [entry field] (in years)
12. How would you define online privacy? [text area]
13. How much would you say you know about online privacy? [radio buttons: *Very little to nothing/A moderate amount/A lot*]
14. How would you define online security? [text area]
15. How much do you know about online security? [radio buttons: *Very little to nothing/A moderate amount/A lot*]
16. Do you think your child always use electronic device(s) securely? (smartphone, desktop, laptop, tablet) [radio buttons: *Yes/No/I'm not sure*]
17. How would you define risky online behavior? [text area]

18. Have your spoken to your child about risky online behaviors? [radio buttons: *Yes/No/I don't remember*]
19. Do you think your child has been a victim of risky online behaviors? [radio buttons: *Yes/No/I'm not sure*]
20. Do you think your child has knowingly engaged in risky or negative online behaviors? [radio buttons: *Yes/No/I'm not sure*]
21. Do you think your child has more, less, or about the same, knowledge of technology as you? [radio buttons: *More/About the same/Less/I'm not sure*]

Semi-Structured Interview Scrip – Parents

1. Tell me about how you spend most of your time online. Do you think you and your child do similar things online?
2. Do you or does someone else monitor and/or limit your child's cell phone use? [If so, who and how? Why?]
3. I see you defined online privacy as [*response from questionnaire Q12*]. How, if at all, do you think online privacy matters in your child's life?
4. Describe good and bad online privacy choice. What are some of the consequences when children your child's age make bad online privacy choices? Have you talked with your child about these choices and these potential consequences?
5. I see you defined online security as [*response from questionnaire Q14*]. How, if at all, do you think online security matters in your child's life?
6. Describe good and bad online security choice. What are some of the consequences when children your child's age make bad online security choices? Have you talked with your child about these choices and these potential consequences?
7. I see you defined an online risk as [*response from questionnaire Q17*]. What sorts of risky choices do you think children your child's age make online? Why do you think children take online risks?
8. What are the most important things you think a child can do to stay private and secure online? What challenges, if any, do you face in helping maintain your child's privacy and security online?
9. Who do you think is most responsible for keeping your child private and secure online?

Appendix C

Example of coding process from first cycle codes through final theming of data

Research Question	First Cycle Sorting Codes	First Cycle Comparison/Discussion Codes	Themed Data
RQ3: How, if at all, do parents influence children's OPS understandings?	• P1a: Describes privacy (including extensions and descriptions from interviewee prior to the question 'do you think online privacy is important?') • P1b: Y1b: Privacy understandings (own or related to child) • Y1a: Describes privacy (including extensions and descriptions from interviewee prior to the question 'do you think online privacy is important?') • Y1b: Privacy understandings (including examples and explanations from interviewee prior to the question 'do you think online privacy is important?) • P2a: Describes online security (including extensions of defi • P2b: Online security understanding (own or related to child; include answers to "how do you know they're secure") • Y2a: Describes online security (including extensions and descriptions from the interviewee prior to the question' do you think online privacy is important?') • Y2b: Online security understanding (including examples and explanations) • P1c: Perception of the role/importance of online privacy to child • P2c: Perception of the role/importance of online security to child • P4d: Reports discussion privacy/security with children as a way to teach/regulate use (include stories) • P4e: Reports physically controlling/monitoring children' devices in some way • P4m: Does not monitor device use/activities	Abbreviated Codes: • Shared knowledge • Fatalistic • Agency/agentive • Believe in privacy • Conflicting beliefs • Consequences • Specific conversations • General "conversations" • Reactive conversation • Consequence-based conversations • Physical control • Device/technology monitoring • Privacy/security connection • Passive monitor • Screen time • Incident-based beliefs • Good kid syndrome • "Open door policy" • Cancel culture • Stranger danger • Deception-as-strategy	Steps: • Compare P1a & P1b with Y1a & Y1b at the dyad level and then cross dyad/grade level • Compare P2a & P2b with Y2a & Y2b at the dyad level and then cross dyad/grade level • Compare P1c & P2c with P4d, P4e, & P4m at the dyad and then cross dyad/grade level; then compare these results with Y1a, Y1b, Y2a & Y2b Results: • Parents who don't find P/S important don't talk about it • Parents of young children rely on parental controls, but don't love them • All monitoring decreases as youth age (within and cross case) • Most parents physically monitor AND have conversations • School is a trusted source • Conversations are reactionary and consequence-centric • Conversations are developmentally-perceived by parents • Higher knowledge = parents who have more conversations (younger) • More specific conversations = higher knowledge (younger)

ImageAlly: A Human-AI Hybrid Approach to Support Blind People in Detecting and Redacting Private Image Content

Zhuohao (Jerry) Zhang[1], Smirity Kaushik[2], JooYoung Seo[2], Haolin Yuan[3]
Sauvik Das[4], Leah Findlater[1], Danna Gurari[5], Abigale Stangl[1], Yang Wang[2]
[1]University of Washington, Seattle [2]University of Illinois at Urbana-Champaign
[3]John Hopkins University [4]Carnegie Mellon University [5]University of Colorado Boulder
{zhuohao, leahkf, astangl@uw.edu}@uw.edu, {smirity2, jseo1005, yvw}@illinois.edu
{hyuan4}@jhu.edu, {sauvik}@cmu.edu , {danna.gurari}@colorado.edu

Abstract

Many people who are blind take and post photos to share about their lives and connect with others. Yet, current technology does not provide blind people with accessible ways to handle when private information is unintentionally captured in their images. To explore the technology design in supporting them with this task, we developed a design probe for blind people — ImageAlly — that employs a human-AI hybrid approach to detect and redact private image content. ImageAlly notifies users when potential private information is detected in their images, using computer vision, and enables them to transfer those images to trusted sighted allies to edit the private content. In an exploratory study with pairs of blind participants and their sighted allies, we found that blind people felt empowered by ImageAlly to prevent privacy leakage in sharing images on social media. They also found other benefits from using ImageAlly, such as potentially improving their relationship with allies and giving allies the awareness of the accessibility challenges they face.

1 Introduction

A challenge for blind people[1] is how to remove private information they unintentionally capture in images they take before sharing the content with others (e.g., personal information on stray screens or pieces of paper, human faces that were not supposed to appear). For example, prior work reported that over 10% of over 40,000 images taken by blind people contained private information [19]. Yet, sharing images is a key way for people to connect with each other, including on social networking services (SNSs) [36]. This challenge on how to preserve private visual information is relevant for the more than 49 million blind people around the world [1,11,29,50,51]

Given the increasing ubiquity and accessibility of mobile devices with built-in cameras, there is a growing potential benefit of developing technology that supports blind users in redacting private information in images. Yet, this capability is not yet available. For instance, a potential workaround is to leverage existing image editing tools to redact private information in images, yet such tools are inaccessible to blind people. That is because such tools require precise hand-eye coordination (e.g., moving the mouse or finger to brush over specific areas). *Our goal is to bridge this gap by empowering blind people with an accessible tool that facilitates the detection and redaction of private content in images they intend to share with others.*

We introduce a new human-AI hybrid approach to enable blind people to avoid unintended privacy-violating disclosures in images they intend to share publicly. We first employ computer vision to provide first-pass prescriptive insights (object recognition and image captioning with associated confidence scores) about image content. Blind users can then decide for themselves whether they consider these identified objects as unnecessarily private or sensitive and whether they want to ask their trusted sighted allies (family members or friends) for targeted editing assistance. If they choose to do so, they can specify how they want the image to be edited (e.g., blurring specific human faces or cropping out certain parts of the image). Our approach was inspired, in part, by what Nissenbaum et al. call *handoff* [40] and what Zhang et al. call an *assistive transfer system* [54]: a system that allows blind people to solicit just-in-time, targeted assistance from a trusted sighted ally to solve an outstanding accessibility challenge. This human-AI hybrid approach also aligns with the emergent perspective of embracing interdependence in assistive technology design [8]. We designed a proof-of-concept system to operationalize and assess this hybrid approach: *ImageAlly*. We aim to use this system as a probe to understand how such tools can be used by blind users and their sighted allies [26].

[1]We use the identity-first language when describing people with visual impairments, guided by the National Federation of the Blind.

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.
USENIX Symposium on Usable Privacy and Security (SOUPS) 2023.
August 6–8, 2023, Anaheim, CA, USA

We conducted a user study to deploy this probe with 20 participants (10 pairs of one blind individual and one sighted ally), following a pilot study (see appendix A.2) with seven participants (four blind people and three sighted allies). The goal of our study was to answer three research questions:

1. How well does our human-AI hybrid approach address blind people's need to identify and redact private information in images they consider sharing online?
2. Given that both the AI-generated insights and the edits generated by human allies provide different levels of information and carry some uncertainty (e.g., AI results can be inaccurate and allies' edits can be subjective), how might blind people use human versus AI assistance? As part of this, we are interested in how they might deal with the uncertainty when deciding whether to solicit targeted editing assistance from trusted allies and share edited images online.
3. How might use of ImageAlly affect the perceived relationship between blind people and their trusted allies, given that prior research on friendsourcing in general [56] and assistive transfer systems in particular [54] suggests that friendsourcing approaches can impact social relationships between friends?

We found that the ImageAlly approach showed promise in supporting blind people in sharing images in a way that aligns with their personal privacy preferences by facilitating the detection and redaction of private content in those images. We also found that our blind participants varied in what they wanted out of ImageAlly. For example, our participants wanted different things out of the AI-powered image screener: some preferred minimal descriptions of image content so they could *efficiently* check for privacy leaks in images they took themselves, while others wanted to use the screener to confirm that their allies appropriately redacted private information in their images. For images processed by allies, we also observed some inconsistencies between sighted allies' editing and blind users' preferences in six out of a total of 20 cases, which were perceived differently by different blind participants and could potentially be avoided by using ImageAlly for a second-time AI screening. Lastly, some participants also believed that their interactions with the blind individuals or sighted allies through ImageAlly have potentially positive impact on their relationships. For example, some sighted ally participants felt that ImageAlly has a positive value in improving their awareness of the challenges that their blind family members or friends faced. They also found ImageAlly useful and felt it could prevent their blind family members or friends from accidentally sharing private information with others.

To summarize, our work makes three main contributions: (1) we introduced and explored the design space of assistive transfer systems for processing images with private information, (2) we designed and implemented a proof-of-concept assistive transfer system, ImageAlly, to serve as a design probe to explore our human-AI hybrid approach in facilitating the detection and redaction of private photo information for blind people, and (3) we conducted a design probe study with both blind people and their sighted allies to answer our research questions and synthesize design insights for assistive transfer systems and other tools designed to improve blind people's exploration and editing of images.

2 Related Work

2.1 Image Sense-making for Blind People

One approach that blind people currently take when they want to make sense of images is to rely on apps and services that use state-of-the-art computer vision techniques to detect objects and caption images. Such tools include Microsoft's Seeing AI [2], Google's Lookout App [16] and other automated image description services [52]. While these services describe image content, their outputs do not offer prescriptive guidance to assist blind people in identifying and obfuscating private information that may be unintentionally captured in those images.

Besides commercially available tools, the development of deep learning models [17, 31, 42] has spurred the increasing use of automated image description models that can assist in image sense-making [5, 25, 53]. Such technologies simplify a wide range of everyday tasks including identifying objects and recognizing familiar faces or facial expressions [4]. Zhao et al. studied how state-of-the-art computer-generated descriptions in Facebook's photo-sharing feature can help blind people improve the photo-sharing experience [55]. Blind people were also found to place a lot of trust in automatically generated captions for visual content on social media (e.g., Twitter) although the caption may diverge from the visual content [37]. Simons et al. studied crowd workers' motivations and challenges for generating image descriptions to develop automated solutions [44]. Finally, Gurari et al. explored the limitations of modern algorithms in captioning images taken by blind people [20]. While ImageAlly is guided by these prior studies, our design probe is novel in understanding blind users' practices of handling private visual content in a human-AI hybrid fashion.

Another approach blind people currently employ to make sense of images is to rely on human intelligence through crowdsourcing or friendsourcing. Blind people sometimes solicit sighted assistance from remote humans to support visual interpretation and visual question answering tasks. This includes relying on remote professional assistance services, such as Aira [3], asking physically proximate allies for direct assistance [51], and soliciting assistance using commercial and research-based crowdsourcing services, including Be My Eyes [41] and VizWiz [10]. Generally, these services provide blind people with remote assistance from sighted allies who, for example, answer questions about their surroundings or provide vocal-guidance on using inaccessible interfaces. However, a limitation of these human-based services is that they do not directly support screening images (or videos) for private content.

2.2 Photo Practices of Blind People

Blind individuals (including teens [9]) take and share photographs for the same reasons that sighted people do [6, 22, 29]. However, when they share photos, identifying possible private and sensitive information inside can be a challenge [45], not to mention processing that information. Researchers have developed many alternatives to assist blind photography. For example, Google Lookout App [16] and iOS AI-based VoiceOver recognition [49] provide text and audio feedback of what is in the camera field of view.

Vázquez et al. proposed to help blind users aim a camera so that they can know for sure what content is inside the frame [47]. Iwamura et al. [28] tried to solve the same problem by introducing a system that uses an omnidirectional camera. Complementing the plethora of prior work around blind photography, we introduce and evaluate the first prototype directly designed to empower blind photographers to avoid inadvertently sharing private/sensitive content captured in their images. Our work builds off of prior work that investigated obfuscation techniques to mitigate privacy leakage in images, such as via blurring, pixelating, inpainting, and avatars [27, 35]. Our ImageAlly system provides sighted allies with obfuscation options and, to our knowledge, our work is the first to explore blind people's experiences of applying obfuscation for privacy-preservation.

3 Design Considerations

We began designing ImageAlly by identifying the challenges blind people experience with sharing images, especially when those images might contain private information. More generally, we identified three design goals for a hybrid human-AI system for blind users to identify and occlude private information in images based on recommendations from prior literature [45, 55]. A co-author of this paper who is blind also informed the design goals we strove towards.

3.1 Identifying Potentially Private and Sensitive Information in Images

Our first goal is to fine-tune state-of-the-art computer vision models to identify potentially private information in an image to facilitate targeted editing or further description by a sighted ally. This goal emerged based on our understanding that despite impressive advances in facial recognition, object detection, optical character recognition (OCR), and image captioning, the identification of what content may be private is highly contextual and personal [4, 45]. Thus, full delegation of this responsibility to AI may be untenable. Such a system would require going beyond simply identifying and captioning the image contents, to recognizing the content in relation to blind people's specific visual privacy concerns in a particular sharing context. In turn, our objective was to fine-tune existing AI models to better identify image content that blind people might consider to be private, e.g. faces and text [45, 55], and then transfer that image content to sighted human allies who can further interpret and redact the private content. We consider such information screening as a first-pass prescriptive insight.

3.2 Redacting Private Information in Images

Our second goal is to source human assistance to redact private information and provide description of their operations for blind people to digest what has been changed in images. For example, blind people may want to crop a certain application window out of a screenshot of their laptop screen, blur out personally identifiable information in a document scan, or blur out children's faces in personal photos. These tasks require an accurate understanding of private visual information and precise hand-eye coordination to act on it, such as moving the mouse/finger to the edge of or over the area that needs to be blurred. Therefore, rather than trying to build an automatic image editing tool using AI models, our objective was to source direct human assistance to help edit the photos. One opportunity to address this goal is to enable a remote ally to directly edit the photos. We consider such a photo transferring and editing as a second-pass human-powered editing.

3.3 Communicating and Verifying Screening Preferences

Our third goal is more of an additional consideration to complement the human-AI hybrid approach. While such a hybrid system of (1) first-pass AI-generated prescriptive insights and (2) second-pass human editing can address our first two design goals, privacy needs vary across individuals and sharing scenarios. For example, the designated ally might recognize some personally identifiable information as private, and not recognize other information as private, and return a photo that does not meet the blind user's expectation (e.g., blurring faces that were not meant to be blurred, or forgetting to crop parts of the image that were meant to be redacted). Given that blind people may not be able to confirm if the edited photo was edited in line with their expectations, a third goal was to provide an accessible way for blind people and their allies to communicate expectations, preferences, and actions. Blind people should have a way to directly state how they expect the photo to be edited.

4 ImageAlly System

Guided by our design goals, we implemented ImageAlly as a design probe [26] on iOS using React Native. In addition to the mobile app, ImageAlly also includes the interface for allies and the backend server. The interface for allies presents the photo sent for redacting, the blind requester's instructions for how they would like the image edited, and an interactive image-editing tool. Next, we describe ImageAlly's interfaces for the blind users and their allies.

4.1 Non-Visual Interface

We designed and developed ImageAlly's non-visual interface (see figure 3) to provide blind users with image descriptions (descriptive screening results) through text (and sound when accessed through a screen reader). To do so, ImageAlly first employed existing libraries and APIs [43] to detect potentially privacy-intrusive information — i.e., faces, pre-selected object categories (e.g., documents, ID cards), and texts that appear in photos based on insights from prior work [19,45] and our blind coauthor's personal experience. This variety of information was collected to assist blind users in their decision-making process while having them ultimately determine if the identified content is private or sensitive, and if so, whether to edit it or leave it as is. Accordingly, in the case that the descriptive screening results indicate that there is potential private information in the image, the interface provides users with a choice to send the photo to their designated ally. As part of this process, the user can specify preferences—by choosing from a list of common options or by typing in their own preferred message for the ally—to indicate how the image is further evaluated for private information.Lastly, users will be asked to select a contact and click a button to send the photo-processing request. Figure 2 summarizes the interaction workflow of ImageAlly. We provide a detailed description of ImageAlly's descriptive screening process in the appendix.

4.2 Visual Interface for The Allies

Once a blind user obtains AI-generated descriptions of potentially private content in images, they may next choose to solicit assistance from allies to redact this information. These sighted allies are solicited through an SMS or email message in which they are provided with a link. The link, in turn, directs the ally to a web interface that presents the photo to be edited, the blind users' corresponding instructions for what information to redact, and a suite of controls to help with redacting private information in images. For example, if the blind user asks the ally to blur out all the text in the photo, the ally can use the built-in tools to blur out the image partly and return the image back to the blind person. Allies also have a text-input box through which they can inform the blind requester of what they did to the photo.

We provided as obfuscation techniques pixelating and blurring using finger-drawing (like an eraser). Note that we use the term "blur" in ImageAlly and throughout the paper as a general term for obfuscation, unless noted otherwise, since we used this term with our study participants to make it easier to understand than with a more technical term such as obfuscation. Of note, prior work has shown that obfuscation techniques such as blurring and pixelating can be ineffective [33–35, 48] or attacked (reversed) via deep learning [39], however, they are still favored by users and viewers [13, 23, 35, 48]. Considering the privacy-utility trade-off and the required effort for obfuscation, we chose *blurring* and *pixelating* in our current design as simple interactions for sighted allies to perform. With that said, ImageAlly could incorporate and work with other current and future improved obfuscation techniques.

5 Study Method

We used ImageAlly as a probe that serves the design goal of inspiring users and researchers to think about new technologies and the social science goal of understanding the needs and desires of users in a real-world setting [26]. Specifically, we sought to gain insights into blind users' perceptions of, preferences towards, and usage of a hybrid human-AI assistive transfer system for identifying and redacting private information in photos they intend to share online. To that end, we conducted an exploratory study of ImageAlly with 20 participants (10 blind people and 10 allies) using an IRB-approved protocol. We asked blind people and one of their sighted allies (a friend or family member, recruited with the blind participant) to use ImageAlly to screen and edit photos from different sources.

After the study, we conducted a comparative analysis on the pictures initially selected by the blind participants and the redacted pictures edited by the allies. This comparative analysis highlighted differences in how participants used ImageAlly, as well as afforded us insight into whether ally-edits aligned with blind participants' preferences. Furthermore, for those ally-edited pictures that did not fully match blind participants' preferences, we followed up with the blind participants and asked them how they felt about and wanted to act on that inconsistency. Together we evaluated how ImageAlly worked in detecting and redacting private image contents and covered cases where ImageAlly did not work perfectly and how blind users would like to handle it.

5.1 Participants

We recruited participants in pairs: one blind user and one sighted ally who the blind user considered a trusted friend or family member. In total, we recruited 10 pairs for 20 total participants: 10 blind participants (referred to as requesters and numbered from R1 to R10), and 10 sighted allies accordingly (referred to as A1 to A10). Also two requesters reported to have hearing impairments. The relationship of the participant pairs varied from friends to family members including mother and daughter, brother and sister, husband and wife. Note that we only recruited requesters that use iPhones.

5.2 Apparatus

We used the ImageAlly design probe to conduct the exploratory lab study. We provided a downloadable link via TestFlight [46] before the study to let the blind users install ImageAlly on their iPhone. For sighted allies, we also designed a simple web interface that contains basic image editing tools including

functions, such as blurring, cropping, and drawing overlay markups (Figure 3, right). We developed the system so sighted allies would not themselves need to install ImageAlly, but would instead receive SMS text or email messages with an embedded link assigned to this photo-editing session.

To simulate different scenarios, we used two image sources in the study. First, we asked each blind user to prepare an image that contains information they consider private. To protect their privacy, we asked them to use outdated information: (1) their room surroundings, (2) selfie or family photos, (3) a screenshot of phone chat history, (4) a received letter, (5) an expired credit card, (6) an expired ID card, (7) a medicine bottle with descriptions, (8) visited webpages. These are the main privacy categories identified in the VizWiz-Priv dataset [19]. We ensured using their photos only for this project.

Second, we asked the blind users to share/re-post an image prepared by our research team on social media. This is to simulate the situation where they share others' visual content. The image was a mobile phone screenshot of a work group's chat history with co-workers' names and avatar profiles.

By using two different sources, we were better able to evaluate ImageAlly by accounting for a broader variety of real world scenarios in which a blind person may consider soliciting assistance, e.g., capture photos, and/or share and repost photos from a second party.

5.3 Procedure

First, we conducted a single session remote study over Zoom. The remote aspect of the study enabled our research team to simulate the likely use-case for ImageAlly, where the requester and the ally are not co-located when the requester might need to use ImageAlly. Upon receiving participants' written consent, we video-recorded all sessions and took detailed notes.

All study sessions lasted about an hour, including a post-study interview to gain insights of requesters' and allies' feedback separately in Zoom breakout rooms. Prior to the study, the participants were told to prepare one photo from the categories mentioned above in 5.1.2, and install the ImageAlly App, which took around 5-10 minutes.

After the study preparation and introduction, the researchers divided the participants into two Zoom breakout rooms to simulate remote collaboration (i.e., they did not need to be physically co-located to use the system). The two researchers who helped conduct the study went into each of the two breakout rooms to guide them through the study, answer their questions, and conduct the exit interviews. After the researchers and the participants settled in different rooms, the researcher in the requester room (referred to as Researcher 1) introduced the tasks and asked the participants questions from a pre-study questionnaire (shown in Appendix A.6) about their experience with photo sharing. The researcher in the ally's room (referred to as Researcher 2) also introduced the tasks and asked the allies questions about their previous experience of receiving requests and providing visual assistance by describing the content of the photos, their concerns about seeing private information from others, and their preference about being contacted by requesters.

After asking both participants about their previous experiences with requesting and/or providing visual assistance, the researchers explained the possible scenarios in which ImageAlly could be used. We asked the requesters to imagine that they were sharing photos across two scenarios that were meant to approximate distinct real-world situations in which blind people may want to share a photo but may harbor concerns about photo content: sharing original photos taken by themselves, and (re-)sharing photos taken by others. Doing so allows us to compare/contrast preferences across different contexts of use — for example, would participants have different privacy concerns when sharing others' vs. their own photos? Would participants want the system configured, and if so, how? To strengthen ecological validity from the ally's perspective, they were instructed that the requests from their friends may come at any time, and that they could do other tasks rather than passively waiting for ImageAlly requests. When their assistance was requested, they would be notified via SMS text message.

After explaining the scenarios, the lab study began. Researcher 1 asked the requester to navigate to the ImageAlly App, go over the instructions in the App, and follow all the prompts from step one to step four (figure 3 left). For the first session, Researcher 1 asked the requesters to use their own photos and answered any question they may have during use of ImageAlly. Within the app, users have the option to send images to allies for editing depending on whether they believe there was private information in images. However, in our study, because the blind participants were instructed to prepare photos with private information prior to the study, most of them (9 out of 10) chose to continue sending the request since there was private information in the images. Only one participant (R1) prepared a selfie which didn't have private information they wanted to blur. However, R1 also chose to continue exploring the full features of ImageAlly. Before they sent the request, the blind participants were prompted to select a contact from their contact book integrated inside ImageAlly. Then they clicked a button to send the request.

After the allies received the message, they were asked to open a link with the web interface of the image editing tool inside. They were also prompted to follow the requesters' stated preferences and crop or blur out certain parts of the photos and provide text description of what they did to the photo. After they were finished, they clicked a button to send back the photo and the description, which can be saved and read by the requesters.

The session was repeated for the second scenario where requesters were asked to forward a photo created by our research team. The photo was sent to the requesters either using an email attachment or a Dropbox download link. Their allies also followed the same process of receiving, reviewing, and editing images. After completing all the tasks,

the researchers conducted an exit interview asking about participants' detailed experiences with the ImageAlly system. The questions include general feedback, suggestions for improvement, Likert-scale questions on system usability, and how such requests might impact the social relationship dynamics between requesters and allies (see Appendix A.7).

After the study sessions, we observed that for some images, there were some inconsistencies between the blind participants' preferences, sighted allies' edited images, and/or their descriptions of how they edited the images. We then analyzed those images and further followed up with the blind participants whose preferences were not fully addressed in the edited images. We asked those participants how they felt about and how they would handle the inconsistency. We report on this post-study analysis in Section 6.4.

5.4 Data Analysis

Upon receiving participant consent, we recorded the study Zoom meeting and logged users' behaviors in the ImageAlly prototype (e.g., blind users' requests and preferences, and how allies edited images) as suggested by Hutchinson et al. [26]. We transcribed the study videos and two members of the research team analyzed the study sessions using thematic analysis [12]. We first individually read and familiarized ourselves with the transcripts. Next, we performed an open coding of sessions independently. We then discussed regularly and eventually converged on the codes and the groupings of codes (i.e., themes) emerged. Since this work is exploratory and our analysis involved the generation of new codes, following guidelines from prior work [38], we did not calculate inter-coder reliability. Example themes include blind participants' prior experience, general feelings about ImageAlly, and how they reacted to the AI and ally generated results about photos. For sighted allies, our analysis covered their impressions about ImageAlly, their willingness and general availability to help blind requesters with images. We also recorded task completion time and their responses to System Usability Scores [15].

We conducted a comparative analysis on the 20 pictures used in the study (10 original pictures selected directly by blind participants, and 10 selected by researchers to be "forwarded" by participants) and the processed version of those pictures edited by allies during the study. Two researchers examined each image manually by independently recording the differences between the picture sent by the blind participant and the picture edited by the ally. We mainly analyzed: (1) What was the image about? (2) What were the blind participant's preferences, how does the edited image look like, and what was the ally's description of how he or she edited the image? (3) What was the difference between the AI screening results of the original image and the edited image?

Then two researchers met online and discussed consensus of these recorded differences. These differences were fairly straightforward to annotate (e.g., whether a person's face has been blurred or not). We mainly focused on gauging the alignment between blind participants' stated preferences for edits and how sighted allies actually edited the pictures.

6 Results

To contextualize our study findings, We first present results that answer our three research questions (from subsection 6.1 to 6.3 accordingly). Finally, we talk about a post-analysis of inconsistent image editing from allies (subsection 6.4).

6.1 Overall Impression and Use of ImageAlly

To answer our first research question about how blind participants and sighted allies felt about ImageAlly's features, we asked participants about their overall impression of ImageAlly. All participants reported that they liked the ImageAlly system but also identified specific pros and cons. To frame their reactions to the system, we next report on our observations of how participants used ImageAlly in the study.

6.1.1 ImageAlly Usage

All 10 blind participants successfully installed the App prior to the study or with a researcher's help during the study, selected the photos they wanted to screen, and received the AI screening results. The photos chose by participants for the study included selfies, family photos, screenshots with personally identifiable information, expired ID cards and credit cards, and document scans. We present details of these photos in table 2 (Appendix A.4).

Each task, from opening the app to receiving the edited photo and saving or sharing it, took between 5 to 15 minutes. We note that the allies were ready to help immediately, which might not always be the case in practice. Task duration depended on several factors such as how familiar both requesters and their allies were with the ImageAlly interface (for the second scenario/task) and whether they asked questions during the task. However, ImageAlly provides an asynchronous way to process photos and it is often not a time-sensitive or urgent task. Most blind participants (9 out of 10) said they were willing to wait for the request to be completed, since they understood that, for instance, the human-editing process could take time. Unlike the first photo task where blind participants used their own images, in the second photo task we asked participants to imagine that they were *"forwarding"* photos from others. All blind participants were able to follow the same steps as the first task.

We also discovered that blind participants used ImageAlly differently across the two scenarios. For their own photos, two blind participants (R2 and R8) took the photos days before the study and thus couldn't locate the photos instantly. Then they used ImageAlly's AI screening function as a confirmation tool to help find the right photo. R2 mentioned that ImageAlly provides a quick and accessible way to confirm whether this photo was the one they wanted to select because it's *"just a couple clicks away."* They already knew roughly what was in

the photo and could quickly recognize the simple keywords or key objects in the AI screening results. For example, R2 found the keyword *"server"* in the AI screening results and immediately recognized that was the right photo she intended to select. In contrast, when "forwarding" others' images, most blind participants used ImageAlly as an exploration tool (as opposed to confirmation) in order to understand the contents of the images because they had no idea.

6.1.2 Blind Participants' Impressions of ImageAlly

All blind participants liked ImageAlly's overall functionality and workflow. They spoke positively about ImageAlly providing: (1) some level of independence and interdependence (e.g., R2 said *"I love the idea, it basically gives me the freedom to do stuff myself, and it's a really great way for my family to assist me"*); (2) more information about photos (e.g., R4 said *"I think it's really cool. I have a lot of experiences with image description but all of them are limited. It's giving me much information"*); and (3) accessible and user-friendly interface (e.g., R3 commented *"I enjoyed it because it's simple to use. It makes sense once I get it and it's pretty user-friendly"*).

Many participants (five blind participants and six ally participants) also pointed out the limitations of ImageAlly. First, while the system allows allies to provide a description of what the photo is about and what they did with the photo, it was still sometimes hard for blind requesters to know what was changed and to trust that the edited photo was free of private information. Two blind participants (out of 10) expressed concern about the edited photos. For instance, R1 mentioned that *"Maybe they missed something or I missed something. Before I share, I want to be confident of what to share."* R1 was worried that perhaps her preference recorded wasn't clear enough for the ally or the ally misinterpreted the message and edited it unexpectedly. However, the other blind participants (8 out of 10) expressed their trust towards their friends and family members in whether they could successfully edit the photo as requested. For instance, R3 said that *"If I choose this friend to send the image, it means I trust them and along with the photo they edited"*. R2 had a similar sentiment: *"I don't need another way (to confirm) because I trust my friend and family"* Note that ImageAlly does not introduce a new trust challenge — even with face-to-face assistance, blind people still face the same challenges with trusting that their ally accurately edited their photo in accordance with their preference. In fact, ImageAlly provides a partial solution to this trust challenge: requesters can run an edited image through the ImageAlly to get some descriptive insight into how an ally's edits changed what was perceivable to the AI, and can just as easily solicit a second opinion from another trusted ally.

Participants also suggested other areas for improvement. For instance, four blind participants said that they wanted to receive notifications when their allies received the request for photo editing to remove the private content, when the allies start working on the request, and when the allies finish checking the photos. They also desired a way to check their allies' availability before they send the request and the option of sending requests to multiple people at the same time when they are unsure if someone is not available.

6.1.3 Allies' Impressions of ImageAlly

All ally participants found the tool useful and easy to use in general. Allies highlighted that ImageAlly could prevent their blind friends or family members from accidentally sharing sensitive information such as credit card information with others. They also felt that the tool readily provides them with ways to help their blind friends or family members. For instance, A8 shared that using ImageAlly would make her *"feel more confident when my husband has to send pieces of info to someone."* She further highlighted that ImageAlly eliminates the need for her to be physically present to help her husband, *"He doesn't necessarily need me right there, he can be in office and me at home and still help him out."*

Some ally participants were even interested in using the blur feature of the tool for their own photos because they were not aware of any other tool that provides the similar blur feature. Our ally participants also offered design suggestions for the tool, such as the ability to zoom into the picture to precisely blur required information. Some allies (A2, A3, A8, A9) also found the instructions provided by the blind requester a bit confusing and hard to interpret. For instance, according to A8, *"Blur my identifiable information is confusing whether it should include only their information or everyone else's too."* Future designs of tools like ImageAlly could allow back-and-forth communication between requesters and helpers.

6.1.4 Perceived Usability of ImageAlly

We also used the System Usability Scale [15] to measure our participants' perceived usability of ImageAlly. Most participants (both requesters and allies) agreed or strongly agreed that ImageAlly was easy to use, had well-integrated and consistent features, and that they would like to use it if ImageAlly is available (figure 1). The calculated SUS scores [15] were 86.25 for blind participants and 84.25 for sighted allies, indicating high usability of ImageAlly.

6.1.5 Privacy Concerns

ImageAlly was designed to process photos with private or sensitive information when blind people wish to share them, either on social media or with friends. However, people may have concerns even when sending them to friends or family members to check. Therefore, we explicitly asked about requesters' concerns of sending photos to an ally. Our blind participants expressed that since ImageAlly allows them to choose the photo and the people they trust to ask for help, they were not concerned. For instance, R6 said *"Now I am confident of what I am going to share."*. Similarly, ally

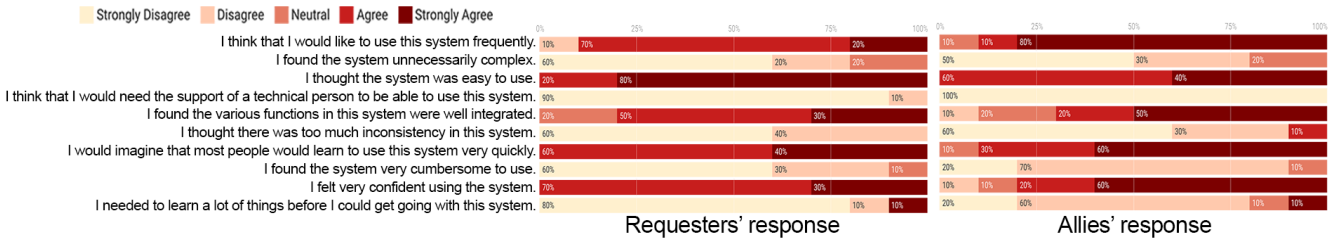


Figure 1: System Usability Scale scores from requesters and allies showing that both blind participants and their allies had a general positive attitude towards ImageAlly's usability.

participants also reported no concerns seeing photos shared by blind participants. However, in some cases, the relationship might impact what photos a blind participant would share with the allies. For example, when A9 was asked, if she had any concerns seeing her mother's private information in the photos, she responded, *"Not an issue so far. My dad is sighted he also checks photos with her."*

6.2 AI and Human-Generated Results

Our second research question focused on how the blind participants perceive and use AI screening and human-processed results. For example, what AI-generated information would be useful for blind people to decide whether to share photos; and how they interpret and use this different information (e.g., type of objects identified, confidence scores of AI results).

6.2.1 Usage of AI-generated Results

Face Number Recognition. ImageAlly provides the number of faces detected in the photo, a feature that most of our blind participants found to be simple and effective. For instance, R8 said that *"When I take objects, I want to know if there are faces I am not aware of, this is quite important."* R3 compared the simple face number recognition with SeeingAI's image exploration functions and thought such detailed and thorough exploration of images in SeeingAI was not necessary when they are taking and checking photos. R5 also pointed out that when they are taking their own photos, they usually *"have a clue of what's going on"* in the photo, and thus simple feedback such as the number of faces recognized is sufficient and more efficient than more detailed descriptors.

Object Detection and Text Extraction. In our study, eight out of 10 photos used by our blind participants had text-based private or sensitive information such as names, addresses, ID numbers (see table 2 in Appendix A.4 for details). All blind participants found that text extraction was particularly important when deciding whether to share photos. In practice, AI-generated outputs have inevitable uncertainty. We were interested in how blind participants would make sense of and make use of the confidence scores of the privacy-relevant AI-generated outputs. For example, how would a confidence score of 50% versus 80% alter a blind user's perception of and trust in the output? We noticed several occasions in the study where the confidence score of a certain object in the blind participant's own photos was low; when that happened, we asked additional questions in the exit interviews about their interpretations of those outputs. We found that while high confidence scores on certain objects (e.g., text of addresses and ID numbers) would unsurprisingly motivate blind people to pay more attention to the photo and transfer the photos to their allies, low confidence scores of any object might also have a similar impact on our blind participants. For instance, R3 said that *"If the accuracy is low, it must be complex in the photos. so I would need it (ImageAlly) even more."* In comparison, R4 interpreted a low score as something unacceptable, *"Low score doesn't make sense to me. They are not as helpful so I would hesitate and even retake the photo before sending it to friends."* Participants tend to believe that low confidence scores often represent complex situations in photos and thus they are more motivated to send requests to friends or family members to process the photos. However, it is hard to define a universal low-score threshold for everyone; standards may from across individuals and contexts. While unpacking the effects of confidence score beyond the scope of our research, understanding uncertainty in AI-generated outputs for systems like ImageAlly appears to be a ripe area for future research.

6.2.2 Usage of Friendsourcing Results

After allies finished editing the photos and sent them back with descriptions, we asked requesters to read through the descriptions and to go over the sharing function either directly with contacts or on social media. Some participants reported that they wanted additional information on the edited photos that were returned, and they preferred to have a reconfirmation of whether the photo had been processed in accordance with their instructions. P5 wanted to have a binary checking result like *"privacy information cleared or not"* using the same AI

algorithm to help them confirm that they can proceed with sharing the photo. However, most participants also expressed their trust in friends and family members and mentioned that they were comfortable with it.

6.2.3 Factors Important for Deciding Whether to Share

We were also interested in what factors might be important when blind participants decide whether to share a photo. Note that because of our study setting (e.g., they already had some idea about the content of the photos they brought to the study), requesters were not making a real decision of whether to share the edited photo or not. However, participants still provided their preferences and thoughts about what factors would be important to them in this decision-making process.

Text cues: The most common factor that has been brought up by nearly all blind participants (9 out of 10) is text cues extracted from photos. Obvious text cues related to personally identifiable information and any number or ID would most likely ring a bell to blind participants and block them from sharing photos without first redacting this information with the assistance of allies. Other types of text cues would also trigger a similar reaction, including (1) extracted text that doesn't make sense (due to the imperfection of algorithms) and (2) long texts that made blind participants realize it's a scanned document that contains lots of information.

Accuracy of screening results: As discussed earlier, we discovered that although low confidence scores in AI-generated outputs can cause confusion, these scores also discourage blind participants from sharing photos and motivate them to use ImageAlly to redact private information.

Description from allies: Allies provided descriptive texts when they finished the requested task. Some blind requesters reported that they wanted to communicate with their allies again when they returned the edited photos. It was either to confirm if the allies had processed some image element specifically that allies didn't specify in the description, or to follow up the request and ask for additional assistance on the same photo. Although blind requesters could source help from the same ally using other social networking services (like directly asking them using Messages or emails), they mentioned that keeping all records on track within a single app is preferred. Unclear or unexpected descriptions from allies that require future attention would discourage our blind participants from sharing photos.

6.3 Perceived Impact on Social Relationships between Blind Individuals and Their Allies

Our third research question focuses on how the use of ImageAlly might the social relationship between blind requesters and their sighted allies. In the exit interview, we asked both user groups about how regular use of a tool like ImageAlly might affect their social relationship. Most participants felt that ImageAlly would bring them closer to the other member of their pair. This result is in line with prior work on assistive transfer systems for solving CAPTCHAs [54], specifically, and for friendsourcing requests generally [56]. For instance, A1 noted that, as a sighted ally and friend, *"I'd have peace of mind that she is not posting anything personal."* R6 felt that ImageAlly makes it more convenient for allies to help because they can do the task remotely on their own devices rather than using blind people's devices: (*"The good thing is that it comes to them"*). R5 mentioned that ImageAlly might afford sighted people better awareness of the experiences of and challenges faced by their blind family members or friends — e.g., the limitations of image captioning systems, which blind people might rely on to make sense of images. R5 added that transferring inaccessible tasks to allies *"enhances their relationship and builds a positive connection"*. While friendsourcing requests in ImageAlly were generally considered beneficial, some participants noted that these requests should be made with good communication and respect for allies' time. For instance, R6 believed that such requests *"will be fine as long as we have good communication about doing things. I just need to be careful and not to rush them."* This result is consistent with the findings from prior work on the potential social costs of friendsourcing [14]. However, their reactions also implied that such social costs could be managed with good communication and awareness of boundaries and limits, echoing findings from prior work on friendsourcing inaccessible CAPTCHA tasks [54].

6.4 Analysis of Ally's Image Editing That Is Inconsistent with Blind Users' Preferences

Upon receiving an editing request with the blind participants' edit preferences, the sighted ally blurred parts of the picture to meet those preferences, and returned the edited picture and a description of the edits back to the blind participant. However, there could be inconsistencies between the blind participants' preferences, the ally-processed pictures, and the sighted allies' descriptions. To assess the frequency with which these inconsistencies might occur, we conducted a post-hoc analysis comparing how well ally-edited images adhered to blind users' stated preferences.

Among the 20 pictures, we found that six (four from blind participants' original pictures and two from researcher "forwarded" pictures) had inconsistencies between blind participants' preferences and how sighted allies processed them (see Table 3 in appendix). We further investigated whether the AI screening algorithms we employed would produce different results for the processed pictures than the original. This differential provided another source of data to assess whether the ally-edited images met the blind participants' preferences. Finally, we followed up with the blind participants in those six cases on how they felt about the inconsistency.

In general, the AI screening algorithms successfully cap-

tured the changes between the original and edited pictures. For example, the information on P7 and P8's cash card and ID card was mostly transcribed by optical character recognition algorithms, which also accurately reflected the remaining information in the processed pictures. The PIN number on P7's card and addresses (although detected in fragmented pieces) on P8's ID card were no long detected in the second-time screening results. In other words, by re-screening ally-edited images, ImageAlly could help blind users confirm whether the private content they wanted to be redacted was effectively redacted by their allies.

In more carefully analyzing the six cases where we observed an inconsistency between blind users' preferences and how allies edited a picture, we observed that a key reason for these inconsistencies was that the blind participants and the ally participants had different interpretations of the former's preferences. For instance, R8's stated preference was to "blur out personally identifiable information" (PII); to that end, their ally blurred out R8's home address but not other PII (e.g., birth date, face) on the State ID card. PII can mean different things to different people, so it is possible that R8's ally had a different interpretation of PII than R8. To simplify the interface, ImageAlly currently offers only pre-defined coarse-grained options for blind users to specify their edit preferences (e.g., remove personal information). However, as we saw in this case, these coarse-grained options may leave too much open for interpretation and could cause inconsistencies between blind users' preferences and how their allies edit their images. One way to help address this issue could be encouraging blind users to provide more specific preferences.

To learn about how blind participants felt about these inconsistencies, we reached out to the blind participants R1, R5, R7, and R8 (from the table above) months after the study. We gave them the accurate descriptions of the processed pictures and asked them whether these pictures met their original expectation and if not, how they felt about this inconsistency. The accurate descriptions were generated objectively by researchers. R1 and R5 expressed that they would not mind the difference. R5 replied that "For me as long as there is some blurring on the image, that is probably fine. It shows that I am trying to protect my privacy, and usually my friends who see my postings would appreciate that." Here, R5 seemed to value more about others' impression of her attitude towards and attempt in enhancing privacy than the completeness of privacy protection.

In comparison, R7 expressed that he wished the sighted ally could "have done a better job at blurring the texts," and also gave other suggestions on sending the preferences of what to do with the pictures more efficiently. R7 said "I wish we could just call them instead of using the App to send the message (preferences in text), then they would probably know what we are talking about." Communicating nuanced preferences through text instructions could be cumbersome; voice-based communications such as phone calls may be more efficient, higher-bandwidth forms of specifying edit preferences. Therefore, complementary communication methods (e.g., voice calls) could be incorporated into the system to make the communication of individual privacy preferences easier.

7 Discussion

In this section, we discuss design implications from user behaviors and our human-AI system, as well as limitations and future work of this work.

7.1 Design Implications

Using ImageAlly as a design probe, our results offer a set of design implications for future private visual content management tools and assistive transfer systems for blind people.

7.1.1 Implications from User Behaviors

First, we discuss a set of implications drawn from our participants' user behaviors, including how they differed by image sources or usage scenarios, how they made sense of objects detected with low confidence scores, and how they had different privacy preferences.

User behavior may differ depending on image sources or use scenarios: Drawing on our findings where blind people behaved differently when checking their own photos versus others' photos, future designers may consider their different behaviors and design image exploring features accordingly. Specifically, They tend to use full features of image exploration when they are exploring other people's photos. In contrast, when checking on photos taken by themselves, they tend to prefer simple descriptions for efficiency.

Object detection with low confidence score might still have value: Blind people often rely much on text descriptions to inform their decisions in sending photo-editing requests and sharing photos. Drawing on our observations, when privacy-related objects are detected by algorithms albeit with low confidence scores, although it will make AI results less credible for decision-making, blind people tend to be more cautious and rely more on transferring the editing tasks, which potentially leads them into paying more attention to private content in their photos. Another related implication is that providing explanations on why the score is low might improve the user experience and give users more confidence when deciding whether to share photos or not.

Individuals can have different privacy preferences: Individuals can have different or even conflicting views on what counts as private content. A profile image might be private to some people but not to others. Moreover, views on what is private and what is not might vary across sharing contexts — for example, if one is attempting to share insurance information with a health provider, then an ID card might be an undue privacy risk. When transferring tasks, it is important to communicate preferences of what private content is and how that content should be processed. Future designs can

explore accessible ways for both parties to communicate and confirm blind users' privacy preferences.

7.1.2 Implications from ImageAlly System

Second, we discuss implications drawn from the ImageAlly prototype system itself, including discussion on usage of human-AI hybrid systems and potential customizable image double-checking for image sense-making.

Using human-AI hybrid systems: AI-based image exploration is often imperfect. Although recent advancement in computer vision has made it much easier and more robust to analyze images [32], it can still lead to confusion and thus human assistance can be useful. In comparison, purely human-based approaches to assist in visual tasks for blind people can be robust and flexible but also slow and expensive [30], which are hard to scale because of people's availability and social cost. Prior research has explored various ways of combining AI and collective human intelligence to tackle accessibility problems, such as using crowdsourcing and computer vision to detect curb ramps [21] or designing crowd-AI cameras to sense the physical world [18]. By exploring human-AI hybrid system's application to image privacy, our design probe also shed light on future designs of assistive transfer systems for blind people in managing private/sensitive visual content. Such hybrid two-layer design can be extended to many other scenarios when AI works at some level but is not perfect.

There is a spectrum of how much AI versus human work should be in this workflow. At one end of the spectrum, AI could do all the work and no humans will be involved. While perfect AI prediction is unlikely in the near future, this is theoretically possible. Previous research suggests that people who are blind tend to have a similar or higher level of privacy concern about sharing their visual content to visual question-answering systems that are powered by humans than powered by AI [45]. If the system completely relies on AI, it is presumably faster and poses less interpersonal privacy risk (since no human counterparties would see the pre-processed image), but the prediction accuracy might not be perfect. At the other end of the spectrum, only humans are involved and there could be multiple human allies involved. Dividing the image-editing task among many allies could reduce the interpersonal privacy risk of crowdsourcing because no single ally would see the entirety of the visual content. However, the task speed would reduce because the completion would depend on the schedule and work from multiple people. We view this AI vs. human design decision as trade-offs between interpersonal privacy, trust, and speed, which future research could explore further. The current, hybrid human-AI design of ImageAlly is already usable and effective, in practice, and is not contingent on any future advances in AI or computer vision.

While our blind participants mentioned in the study that ImageAlly gave them the ability to do things on their own first, the assistive transfer system approach still relies on the interdependence between blind participants and their allies. Interdependence is considered valuable in assistive technologies [8]. ImageAlly does not necessarily change the fact that blind people might seek help from allies. Instead, it provides an integrated way for blind people to transfer the photo screening task with autonomy, and foreshadows future research on improving the accessibility of screening and editing photos for blind people when they have the needs. However, social support is not always appropriate or desired, and thus there are likely limits to its use, such as the social cost of asking for help. While our results suggest that the usage of ImageAlly could actually improve the social relationship between blind requesters and sighted allies, future research could further examine the cost of social support in such systems.

Consider customizable image double-check: Some blind participants wanted to be able to check the edited photos after receiving them back from allies. Specifically, there was a desire to compare the AI screening results before and after their allies' editing. This suggests the option of double checking the ally edited photos and highlighting differences before and after human editing. For instance, AI can be applied again to the ally edited photos and can simply say for instance "the human faces are no longer present." Furthermore, another implication we drew from participants' responses is that there is value in making the double check process customizable. For example, requesters can set a rule like "Face on the right side should not be detectable" and thus, both allies and the algorithm would have a clear metric of what to detect and edit.

7.2 Limitations and Future Work

7.2.1 Limitation of The Lab Study

ImageAlly has some limitations when deployed in our lab study. For example, we found that requesters felt the system did not provide sufficient notifications to them when allies receive the message and start working on it. Participants reported that they prefer to have a way to know if their allies started processing the photos so that they have a better sense of whether to send requests to another ally. Another limitation was that requesters need to select a contact to send the request. Participants wanted more flexible ways to select one or more contacts when they needed to, like maintaining a commonly used friend list. We also did not have a large sample size. It is challenging to recruit blind participants, and it was more so in our study because every session requires a pair of a blind person and an ally. However, our sample size is on par with other privacy/security user studies focusing on blind people [7, 24, 54].

Another limitation stems from the controlled nature of our lab study. We chose to conduct an exploratory lab study rather than a field deployment because we were at an exploratory stage of designing such hybrid human-AI system and we want to obtain rich and qualitative data on users' perceptions of and reactions to ImageAlly as a probe. Also, field deployment is

more appropriate in a later stage of the iterative design process. In the meantime, we also recognize that some study settings like asking participants to imagine how ImageAlly would affect their relationship are limited and can only be answered in a lab study but in a field deployment. We consider this as a promising future work. As a result, the images used in our lab study are limited in representing the types of photos blind people might share in real lives. Additionally, although we told allies that requests from their blind friends or family members may come at any time, and that they could do whatever they pleased in the meanwhile rather than waiting for ImageAlly requests, allies still put themselves in a lab study situation and were always available when requesters sent requests. However, in practice, allies might not always be available at the time when blind people need help. However, compared to prior work on transferring CAPTCHA tasks (which usually expire in 2 minutes) [54], photo screening is often less urgent and thus is an asynchronous task, which can also be sent to more than one ally at the same time. This could help scale ImageAlly since ally's availability is less of a concern.

7.2.2 Limitations of The ImageAlly Prototype

ImageAlly was implemented as a proof-of-concept design probe rather than as a full-fledged production-ready system, and its current implementation is not bulletproof for privacy and security. While blind participants were all comfortable with choosing a trusted ally to deal with their photos, the system might be exploited by malicious attackers. For example, since requests were sent to allies using URLs via SMS text messages (it will come from a phone number used by ImageAlly), it might be intercepted by malicious third parties. In addition, attackers might send malicious requests to unsuspecting allies and get them to edit photos for free. Since ImageAlly was built as a friendsourcing system, requesters and allies should already have a trustworthy relationship. Several strategies could help mitigate the above privacy/security risks, including, for example, requiring registration and authentication on both sides, building a trustable contact list (whitelist), setting request quotas per day, and each party sending a separate confirmation text message to the other party directly using their own phone number. Apart from the security risks of using ImageAlly's transferring feature, there are also privacy and security implications of using 3rd party APIs. This risk could be mitigated by avoiding using 3rd party commercial APIs and developing proprietary machine learning models based on datasets like VizWiz. These are implementation-specific trade-offs, and not fundamental risks imposed by the system design.

Another limitation came from the fact that ImageAlly was built on existing computer vision models to detect objects in visual content. Although our intention was to provide users with full agency and control of their own visual data by listing all possible objects detected to empower their image editing/sharing decision-making, existing object detection algorithms can have false negative errors (e.g., a card on the table was not detected because it was far away from the camera). We consider such inaccuracies as motivations for future AI solutions. Also, image analysis contains a variety of means beyond what we proposed in section 4.1, we consider studying what image analysis is necessary and efficient for blind users to make sense of pictures as new challenges for future work.

7.2.3 Future Work

There are several areas of future work that this work opens. One potential avenue is the exploration of the critical role that allies play in our human-AI system. While our current research scope focuses on blind users, It is necessary to study the overall satisfaction of allies and how they are impacted with unintended social tensions. We can also explore how such system can encourage allies to engage in visual tasks related to blind individuals' privacy. Another potential direction is to use more advanced computer vision techniques to better study the dynamics between human and machine intelligence with a more robust and reliable system. Furthermore, future research can also study additional use cases beyond sharing on social media, as ImageAlly could be used in virtually any case where blind users want to send or share visual content with another party (e.g., uploading an image for registration or reimbursement).

8 Conclusion

To assist blind people in detecting and redacting private content in photos that they might consider sharing online, we designed and implemented a proof-of-concept probe — ImageAlly. ImageAlly employs a hybrid human-AI workflow that affords blind users AI-generated insights about potential private content in images, and then facilitates the solicitation of targeted editing assistance from trusted allies. Through an exploratory lab study with recruited pairs of blind people and their sighted allies, we found that both parties liked ImageAlly. We also found that blind users preferred coarse, minimal descriptors for their own photos (e.g., number of faces detected) but more fine-grained descriptors on others' photos from the AI-generated screening results. Furthermore, our results suggest that use of assistive transfer systems like ImageAlly has the potential to strengthen the relationship between blind requesters and their sighted allies. ImageAlly could also increase sighted allies' awareness of the challenges faced by their blind friends and family members.

9 Acknowledgements

We thank our participants for their contributions and sharing their insights. This research was in part supported by the National Science Foundation (NSF) grants #2126314, #2028387, #2125925, and #2148080 and the CRA CIFellows program.

References

[1] Tousif Ahmed, Roberto Hoyle, Kay Connelly, David Crandall, and Apu Kapadia. Privacy Concerns and Behaviors of People with Visual Impairments. In *CHI2015*, pages 3523–3532, 2015.

[2] Microsoft Seeing AI. https://apps.apple.com/us/app/seeing-ai/id999062298, Accessed: 2022-01-31.

[3] Aira. https://aira.io/, Accessed: 2022-01-31.

[4] Taslima Akter, Bryan Dosono, Tousif Ahmed, Apu Kapadia, and Bryan Semaan. " i am uncomfortable sharing what i can't see": Privacy concerns of the visually impaired with camera based assistive applications. In *29th {USENIX} Security Symposium ({USENIX} Security 20)*, 2020.

[5] Jyoti Aneja, Aditya Deshpande, and Alexander G Schwing. Convolutional image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5561–5570, 2018.

[6] Giulia Barbareschi, Catherine Holloway, Katherine Arnold, Grace Magomere, Wycliffe Ambeyi Wetende, Gabriel Ngare, and Joyce Olenja. The social network: How people with visual impairment use mobile phones in kibera, kenya. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2020.

[7] Nata Barbosa, Jordan Hayes, and Yang Wang. UniPass: Design and Evaluation of A Smart Device-Based Password Manager for Visually Impaired Users. In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp 2016)*, 2016.

[8] Cynthia L Bennett, Erin Brady, and Stacy M Branham. Interdependence as a frame for assistive technology research and design. In *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility*, pages 161–173, 2018.

[9] Cynthia L Bennett, Jane E, Martez E Mott, Edward Cutrell, and Meredith Ringel Morris. How teens with visual impairments take, edit, and share photos on social media. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–12, 2018.

[10] Jeffrey P. Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C. Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, and others. VizWiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, pages 333–342. ACM, 2010.

[11] Rupert RA Bourne, Jaimie Adelson, Seth Flaxman, Paul Briant, Michele Bottone, Theo Vos, Kovin Naidoo, Tasanee Braithwaite, Maria Cicinelli, Jost Jonas, et al. Global prevalence of blindness and distance and near vision impairment in 2020: progress towards the vision 2020 targets and what the future holds. *Investigative Ophthalmology & Visual Science*, 61(7):2317–2317, 2020.

[12] Richard E Boyatzis. *Transforming qualitative information: Thematic analysis and code development*. sage, 1998.

[13] Michael Boyle, Christopher Edwards, and Saul Greenberg. The effects of filtered video on awareness and privacy. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work*, pages 1–10, 2000.

[14] Erin L Brady, Yu Zhong, Meredith Ringel Morris, and Jeffrey P Bigham. Investigating the appropriateness of social network question asking as a resource for blind users. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 1225–1236, 2013.

[15] John Brooke. SUS-A quick and dirty usability scale. *Usability evaluation in industry*, 189(194):4–7, 1996.

[16] Lookout by google. https://play.google.com/store/apps/details?id=com.google.android.apps.accessibility.reveal&hl=en_US&gl=US, Accessed: 2022-01-31.

[17] Arianna Dagliati, Simone Marini, Lucia Sacchi, Giulia Cogni, Marsida Teliti, Valentina Tibollo, Pasquale De Cata, Luca Chiovato, and Riccardo Bellazzi. Machine learning methods to predict diabetes complications. *Journal of diabetes science and technology*, 12(2):295–302, 2018.

[18] Anhong Guo, Anuraag Jain, Shomiron Ghose, Gierad Laput, Chris Harrison, and Jeffrey P Bigham. Crowd-ai camera sensing in the real world. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(3):1–20, 2018.

[19] Danna Gurari, Qing Li, Chi Lin, Yinan Zhao, Anhong Guo, Abigale Stangl, and Jeffrey P Bigham. Vizwiz-priv: a dataset for recognizing the presence and purpose of private visual information in images taken by blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 939–948, 2019.

[20] Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. Captioning images taken by people who are blind. In *European Conference on Computer Vision*, pages 417–434. Springer, 2020.

[21] Kotaro Hara, Jin Sun, Robert Moore, David Jacobs, and Jon Froehlich. Tohme: detecting curb ramps in google street view using crowdsourcing, computer vision, and machine learning. In *Proceedings of the 27th annual ACM symposium on User interface software and technology*, pages 189–204, 2014.

[22] Susumu Harada, Daisuke Sato, Dustin W Adams, Sri Kurniawan, Hironobu Takagi, and Chieko Asakawa. Accessible photo album: enhancing the photo sharing experience for people with visual impairment. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2127–2136, 2013.

[23] Rakibul Hasan, Eman Hassan, Yifang Li, Kelly Caine, David J Crandall, Roberto Hoyle, and Apu Kapadia. Viewer experience of obscuring scene elements in photos to enhance privacy. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2018.

[24] Jordan Hayes, Smirity Kaushik, Charlotte Emily Price, and Yang Wang. Cooperative Privacy and Security: Learning from People with Visual Impairments and Their Allies. In *Proceedings of Fifteenth Symposium on Usable Privacy and Security (SOUPS 2019)*, 2019.

[25] Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. Image captioning: Transforming objects into words. *arXiv preprint arXiv:1906.05963*, 2019.

[26] Hilary Hutchinson, Wendy Mackay, Bo Westerlund, Benjamin B Bederson, Allison Druin, Catherine Plaisant, Michel Beaudouin-Lafon, Stéphane Conversy, Helen Evans, Heiko Hansen, et al. Technology probes: inspiring design for and with families. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 17–24, 2003.

[27] Panagiotis Ilia, Iasonas Polakis, Elias Athanasopoulos, Federico Maggi, and Sotiris Ioannidis. Face/off: Preventing privacy leakage from photos in social networks. In *Proceedings of the 22nd ACM SIGSAC Conference on computer and communications security*, pages 781–792, 2015.

[28] Masakazu Iwamura, Naoki Hirabayashi, Zheng Cheng, Kazunori Minatani, and Koichi Kise. Visphoto: Photography for people with visual impairment as post-production of omni-directional camera image. page 1–9. ACM, Apr 2020.

[29] Chandrika Jayant, Hanjie Ji, Samuel White, and Jeffrey P Bigham. Supporting blind photography. In *The proceedings of the 13th international ACM SIGACCESS conference on Computers and accessibility*, pages 203–210, 2011.

[30] Ece Kamar. Directions in hybrid intelligence: Complementing ai systems with human intelligence. In *IJCAI*, pages 4070–4073, 2016.

[31] Ioannis Kavakiotis, Olga Tsave, Athanasios Salifoglou, Nicos Maglaveras, Ioannis Vlahavas, and Ioanna Chouvarda. Machine learning and data mining methods in diabetes research. *Computational and structural biotechnology journal*, 15:104–116, 2017.

[32] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023.

[33] Pavel Korshunov, Andrea Melle, Jean-Luc Dugelay, and Touradj Ebrahimi. Framework for objective evaluation of privacy filters. In *Applications of Digital Image Processing XXXVI*, volume 8856, pages 265–276. SPIE, 2013.

[34] Yifang Li, Nishant Vishwamitra, Hongxin Hu, Bart P Knijnenburg, and Kelly Caine. Effectiveness and users' experience of face blurring as a privacy protection for sharing photos via online social networks. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 61, pages 803–807. SAGE Publications Sage CA: Los Angeles, CA, 2017.

[35] Yifang Li, Nishant Vishwamitra, Bart P Knijnenburg, Hongxin Hu, and Kelly Caine. Effectiveness and users' experience of obfuscation as a privacy-enhancing technology for sharing photos. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):1–24, 2017.

[36] Yiyi Li and Ying Xie. Is a picture worth a thousand words? an empirical study of image content and social media engagement. *Journal of Marketing Research*, 57(1):1–19, 2020.

[37] Haley MacLeod, Cynthia L Bennett, Meredith Ringel Morris, and Edward Cutrell. Understanding blind people's experiences with computer-generated captions of social media images. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 5988–5999, 2017.

[38] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for cscw and hci practice. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–23, 2019.

[39] Richard McPherson, Reza Shokri, and Vitaly Shmatikov. Defeating image obfuscation with deep learning. *arXiv preprint arXiv:1609.00408*, 2016.

[40] Deirdre K Mulligan and Helen Nissenbaum. The concept of handoff as a model for ethical analysis and design. In *The Oxford Handbook of Ethics of AI*, page 233. Oxford University Press, 2020.

[41] Be my eyes - see the world together. https://www.bemyeyes.com/. Accessed: 2022-01-31.

[42] Pedro J Navarro, Carlos Fernandez, Raul Borraz, and Diego Alonso. A machine learning approach to pedestrian detection for autonomous vehicles using high-definition 3d range data. *Sensors*, 17(1):18, 2017.

[43] Cognitive services — APIs for AI developers. https://azure.microsoft.com/en-us/services/cognitive-services/, Accessed: 2022-01-31.

[44] Rachel N Simons, Danna Gurari, and Kenneth R Fleischmann. " i hope this is helpful" understanding crowdworkers' challenges and motivations for an image description task. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–26, 2020.

[45] Abigale Stangl, Kristina Shiroma, Bo Xie, Kenneth R Fleischmann, and Danna Gurari. Visual content considered private by people who are blind. In *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility*, pages 1–12, 2020.

[46] Testflight. https://developer.apple.com/testflight/, Accessed: 2022-01-31.

[47] Marynel Vázquez and Aaron Steinfeld. Helping visually impaired users properly aim a camera. In *Proceedings of the 14th international ACM SIGACCESS conference on Computers and accessibility*, pages 95–102, 2012.

[48] Nishant Vishwamitra, Bart Knijnenburg, Hongxin Hu, Yifang P Kelly Caine, et al. Blur vs. block: Investigating the effectiveness of privacy-enhancing obfuscation for images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 39–47, 2017.

[49] Accessibility - Vision. https://www.apple.com/accessibility/vision/, Accessed: 2022-01-31.

[50] Violeta Voykinska, Shiri Azenkot, Shaomei Wu, and Gilly Leshed. How blind people interact with visual content on social networking services. In *Proceedings of the 19th acm conference on computer-supported cooperative work & social computing*, pages 1584–1595, 2016.

[51] Shaomei Wu and Lada A Adamic. Visually impaired users on an online social network. In *Proceedings of the sigchi conference on human factors in computing systems*, pages 3133–3142, 2014.

[52] Shaomei Wu, Jeffrey Wieland, Omid Farivar, and Julie Schiller. Automatic alt-text: Computer-generated image descriptions for blind users on a social network service. page 1180–1192. ACM, Feb 2017.

[53] Zhongliang Yang, Yu-Jin Zhang, Sadaqat ur Rehman, and Yongfeng Huang. Image captioning with object detection and localization. In *International Conference on Image and Graphics*, pages 109–118. Springer, 2017.

[54] Zhuohao Zhang, Zhilin Zhang, Haolin Yuan, Natã M Barbosa, Sauvik Das, and Yang Wang. Webally: Making visual task-based captchas transferable for people with visual impairments. In *Seventeenth Symposium on Usable Privacy and Security ({SOUPS} 2021)*, pages 281–298, 2021.

[55] Yuhang Zhao, Shaomei Wu, Lindsay Reynolds, and Shiri Azenkot. The effect of computer-generated descriptions on photo-sharing experiences of people with visual impairments. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):1–22, 2017.

[56] Haiyi Zhu, Sauvik Das, Yiqun Cao, Shuang Yu, Aniket Kittur, and Robert Kraut. A market in your social network: The effects of extrinsic rewards on friendsourcing and relationships. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 598–609, 2016.

A Appendix

A.1 Non-Visual Interface Details

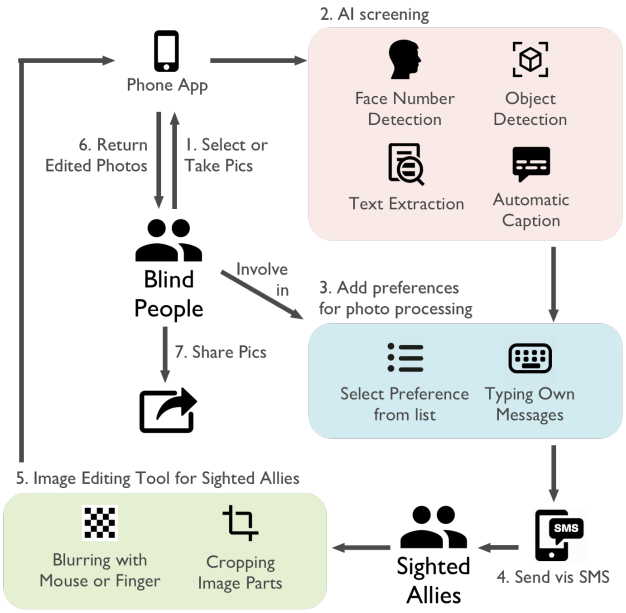


Figure 2: Workflow of ImageAlly: Blind users select or take a picture from the App, uses AI to screen the photo, and read through the results. Then they will be prompted to add a preference of what to do with the photo and choose whether to send it via SMS messages to sighted allies. If they choose to send a request, allies will use an interactive image editing tool to blur or crop out image parts and send it back with descriptions of what they did. Blind people will then have the option to share it with others or on social media.

A.1.1 Detection of the Quantity of Faces:

Whether the number of faces matches blind people's own expectation is an important measurement of whether the image has unnecessary private or sensitive information. For example, if a blind user is trying to take a selfie, a family photo, or photos of scenery, they may have specific expectations of how many faces should appear in the photo. If the number of detected faces diverges from this expectation, the blind user may elect for further screening and processing of the photo.

A.1.2 Detection of Related Objects:

State-of-the-art computer vision models can identify objects in images and describe these objects in natural language. We leverage commercial APIs [43] of state-of-the-art models to provide both object category names and detection confidence scores to help blind users decide whether to transfer the images to human allies for additional processing. For example, if an ID card is detected in the photo, ImageAlly will present the user with a prompt akin to the following: "We have detected the following objects in the picture, together with a percentage number showing how confident we are for each detected object: Text (72% sure), Card (81% sure)." Note that the categories are provided by the Microsoft Azure Object Detection API [43] and therefore we are unable to get a full category list of objects to be detected. We consider this out of our scope because our ImageAlly design is intended to be able to generalize for use with other current and future improved object detection models.

A.1.3 Detection of Related Texts:

If text is detected, we then employ a commercial optical character recognition (OCR) API [43] to extract the text and present it to the user.

A.1.4 Automatic Captioning:

We also use commercial neural image captioning APIs to generate a caption [43], along with an associated confidence score, to help users generally and broadly understand the broad strokes of what is captured in the image. An example caption: "Additionally, we are 93.27% sure that this picture can be generally described as: graphical user interface, text, application, chat or text message."

A.1.5 Detection of Adult Content:

Using Microsoft's visual feature APIs [43], we also added adult content detection.

Note that we are providing as many image analysis result as possible for screening processes. With the development of image analysis techniques, this subsection can go longer as needed. We are using these detection categories to provide as an example in designing ImageAlly probe. Furthermore, the use of AI in ImageAlly is only to provide descriptions of *potentially* private content. It is ultimately up to the blind user to determine if they deem content to be private and so if it should be edited or left alone.

A.2 Pilot Study

Prior to the main study, we conducted a series of pilot sessions with seven blind participants to improve the workflow and accessibility of ImageAlly. We conducted these pilots similar to our full study with our ImageAlly prototype. Specifically, we did four pilot study sessions with seven participants using a task-based usability test.

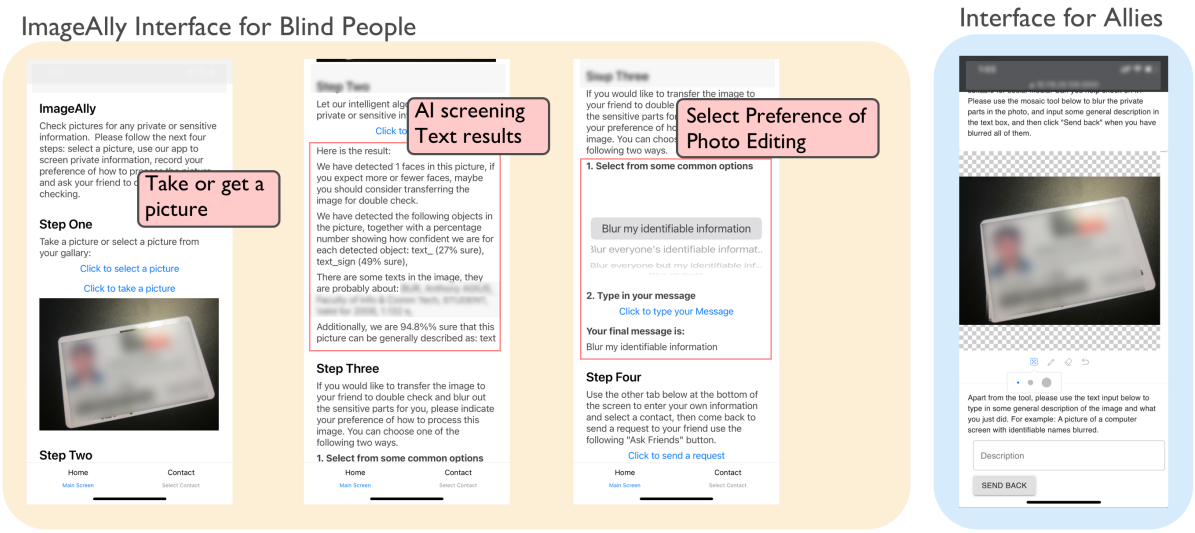


Figure 3: Interfaces of ImageAlly. Left side, UI for blind users: (1) select or take an image, (2) Using images selected in Step one, screen the image using AI algorithms (Note that the detected result of the ID card in Step one is considered as "text" and "text-sign" (hand-written text sign) because the ID card mostly contained text. The detection result is limited to existing commercial APIs, which will be discussed in section 7.2.2), and (3) set image editing preferences and then send a request to allies (The common preference options listed here include blurring out the requester's or everyone's identifiable information or faces. Requesters can also indicate their own preferences). (b) UI for allies: Description of the task, together with AI screening results and a text input for describing their actions.

A.2.1 Pilot Study Method

The first session included one blind participant and one of our researchers acted as the ally upon the participant's consent. The remaining three pilot sessions included one blind participant and one of their sighted friends or family members as an ally. We used our initial ImageAlly prototype as the apparatus for the study, and prepared a image for the participants to use as the material. The image was a mobile screenshot of a work group chat, containing private information like co-workers' names, work content, and their profile avatar images. During each study session, the participants were divided into two Zoom breakout rooms so that they could focus on testing ImageAlly without talking to each other and causing disturbance. First the researchers introduced the study and the task, followed by a series of interview questions about their previous experience of receiving or providing visual assistance regarding photos. Then researchers asked the participants to use ImageAlly to process the prepared photo before they share it to a third party. As the procedure of using ImageAlly in pilot study is identical to the procedure of formal study evaluation, we present in Section 5.3. We recorded and transcribed the data for each session. Two researchers developed themes from the transcripts using thematic analysis. Then researchers met online to discuss how participants' feedback informed improvement of ImageAlly and iterated the system accordingly.

A.2.2 Pilot Study Results

Below we summarize what we learned from the pilot study and the main changes we made accordingly to the system.

Improve interface accessibility. Pilot study participants reported that ImageAlly's interface could be made more accessible. For example, participants mentioned that the navigation inside the ImageAlly probe could be improved with heading and page-based navigation. They also suggested using prompts and confirmations more often when they or their allies finish a certain step. Specifically, they wanted to receive notifications when their allies received the request and started working on it. Based on this feedback, we added heading navigation to the prototype and added notifications using accessible pop-up alerts each time a user completes a step.

Present AI results to allies. Participants also suggested that their allies should view the same screening results from AI as they themselves did. By presenting the AI results to allies, they could understand what their blind friends are getting and can then make better decisions when they edit the photos. For example, there might be privacy-related content in the photo that was not detected by AI, or inaccurate AI predicted results that blind users obtain and use to make decisions. Therefore, we present the AI results to allies when they receive the image processing request.

Improve asynchronous communications. In our original prototype, the communication happened in an unidirectional

way. Only blind people were able to articulate preferences for what they like their ally to do with their photos. However, participants suggested that allies should also be able to respond with what they did to the photos. Based on this, we added a bidirectional communication channel to allow allies to describe the edits they made to blind requesters.

A.3 Participants Biographics

We provide participants' biographics here in table 1.

A.4 Photos Used by Participants

We provide general descriptions of photos (table 2) that contained (outdated) private information of blind participants.

A.5 Comparison Between Inconsistent Image Editing

We provide a comprehensive analysis result of comparison between inconsistent image editing results transferred back from sighted allies here in table 3.

A.6 Pre-study Questions

A.6.1 Questions for Blind Requesters

- Do you take photos? When and how? What kind of photos?
- Do you share your photos with others? When and how? What kind of photos?
- Do you share your photos on social media? When and how? What kind of photos?
- Do you edit your photos before you share? What do you edit photos for and how do you do that?
- How do you decide what photos to share?
- What's usually in the picture parts where you want to edit? How do you usually do that?

A.6.2 Questions for Sighted Allies

- Did your friend/family member (blind or low vision) ever consult you about photos they take? Like whether the image contains the right content, whether the figure looks good, whether there's private information that should be cropped out or blurred out?
- Do you have any concerns seeing their private information in the photos if there's any?
- What's your preference for how to be contacted by the requester? Like text messages, email, DMs from social media App, etc.

A.7 Exit Interview Questions

A.7.1 Questions for Blind Requesters

- Please give us a general impression of the idea and process
- To recap the screening results, how do you interpret the AI outputs? What's useful? What's not useful?
- What do you expect to see more in the AI result for better decision-making on whether to share it to public?
- Have you tried other tools that help you recognize the image contents? What kind of information helps you decide whether there is private or sensitive information?
- How do you make use of the preference recording function (step3)?
- How do you make use of the returned image from friends? Do they meet your expectations?
- Would using this app influence your relationship with [the other party]? How would it potentially affect the relationship in any positive or negative way?
- Do you have any more suggestions?
- Please rate from 1-5 (strongly disagree to strongly agree) for the following statements
  - I think that I would like to use this system frequently.
  - I found the system unnecessarily complex.
  - I thought the system was easy to use.
  - I think that I would need the support of a technical person to be able to use this system.
  - I found the various functions in this system were well integrated.
  - I thought there was too much inconsistency in this system.
  - I would imagine that most people would learn to use this system very quickly.
  - I found the system very cumbersome to use.
  - I felt very confident using the system.

A.7.2 Questions for Sighted Allies

- Please give us a general impression of the idea and process
- How do you think about the instruction by the requester? Is it helpful?
- How do you think about the image editing function?
- How do you think about using SMS text to transfer the request?

Table 1: Blind participants' demographics, including their age group, gender identity, self-described disability, their allies' gender and relationships (allies' relationship to requesters).

Requester	Age	Gender	Self-Described Disability	Ally	Ally Age	Ally Gender	Relationship
R1	35-44	Female	Blind	A1	35-44	Male	Friend
R2	25-34	Female	Blind	A2	55-64	Female	Mother
R3	25-34	Female	Blind	A3	25-34	Male	Friend
R4	18-24	Male	Blind and Hearing Impairments	A4	18-24	Female	Partner
R5	45-54	Female	Blind	A5	18-24	Female	Daughter
R6	55-64	Female	Blind and Hearing Impairments	A6	55-64	Female	Friend
R7	55-64	Male	Blind	A7	55-64	Female	Sister
R8	35-44	Male	Blind	A8	35-44	Female	Friend
R9	25-34	Male	Blind	A9	25-34	Female	Friend
R10	18-24	Male	Blind	A10	18-24	Male	Brother

Table 2: Photos created and used by blind participants

Participant ID	Photo they chose
R1	A man in front of a birthday cake
R2	Mobile phone screenshot with server port and password
R3	Mobile phone screenshot with calendar invite and personal information
R4	Room surroundings
R5	Medical bottle with prescriptions
R6	Insurance document on the table
R7	Expired cash card on the table
R8	Expired state ID card on the table
R9	Transaction screenshot with transaction ID and part of bank account number
R10	ID card on the table

- How do you make use of the text input as a way to inform requesters about what you edited?
- Do you have any more suggestions?
- Please rate from 1-5 (strongly disagree to strongly agree) for the following statements
  - I think that I would like to use this system frequently.
  - I found the system unnecessarily complex.
  - I thought the system was easy to use.
  - I think that I would need the support of a technical person to be able to use this system.
  - I found the various functions in this system were well integrated.
  - I thought there was too much inconsistency in this system.
  - I would imagine that most people would learn to use this system very quickly.
  - I found the system very cumbersome to use.
  - I felt very confident using the system.

PID	Which Picture	Criteria	Description or Quote
R1	Original	Picture Abstract	A man smiling in front of a birthday cake with candles on it
		Blind User Preference	"Blur out the cake"
		Processed Image	The candles on the cake were blurred
		Ally's description	"I blurred out the cake"
		AI Screening Difference	Candle is no longer detected in the processed image, but cake still is
R5	Original	Picture Abstract	A medicine bottle with prescriptions on it, including name, tablet size, prescription, ID, and date
		Blind User Preference	"Blur out prescription information"
		Processed Image	The name, tablet size, ID and date were blurred, but the prescription including the medical condition was not blurred
		Ally's description	"I blurred out the prescription"
		AI Screening Difference	Texts are no longer detected in the processed image
R7	Original	Picture Abstract	A picture of a cash card's back, including security information of card number and PIN
		Blind User Preference	"Blur out security information of this card"
		Processed Image	The PIN was blurred out, but the card number is not
		Ally's description	"I erased the password"
		AI Screening Difference	Text changed from instructions and card number and password to instructions and card number only
R8	Original	Picture Abstract	A state ID card's front page, including name, date of birth, address, biometrics, and face picture
		Blind User Preference	"Blur out personal identifiable information"
		Processed Image	The address was blurred out, but the rest of the information was not
		Ally's description	"I blurred your address"
		AI Screening Difference	Text of address is no longer detected in the processed image
R4	Forwarded	Picture Abstract	Mobile screenshot of a work group chat history with co-workers' names and profile pictures
		Blind User Preference	"Blur out personal identifiable information"
		Processed Image	Only the names were blurred, the profile pictures still remain in sight
		Ally's description	"I blurred their information"
		AI Screening Difference	Text of names were no longer detected
R6	Forwarded	Picture Abstract	Mobile screenshot of a work group chat history with co-workers' names and profile pictures
		Blind User Preference	"Blur out colleagues' info"
		Processed Image	Only the profile pictures were blurred, the names still remain in sight
		Ally's description	"I blurred out their faces"
		AI Screening Difference	Face number changed to 0

Table 3: Comparison between the pictures that contain inconsistent editing with ally processed pictures, including (1) what this picture was about, (2) what was the blind participant's preference for processing the picture, (3) what was the processed picture like, (4) what was the described by allies, and (5) what were the differences between AI screening results for the two pictures.

Evaluating the Impact of Community Oversight for Managing Mobile Privacy and Security

Mamtaj Akter
Vanderbilt University

Madiha Tabassum
Northeastern University

Nazmus Sakib Miazi
Northeastern University

Leena Alghamdi
University of Central Florida

Jess Kropczynski
University of Cincinnati

Pamela J. Wisniewski
Vanderbilt University

Heather Lipford
University of North Carolina, Charlotte

Abstract

Mobile privacy and security can be a collaborative process where individuals seek advice and help from their trusted communities. To support such collective privacy and security management, we developed a mobile app for Community Oversight of Privacy and Security ("CO-oPS") that allows community members to review one another's apps installed and permissions granted to provide feedback. We conducted a four-week-long field study with 22 communities (101 participants) of friends, families, or co-workers who installed the CO-oPS app on their phones. Measures of transparency, trust, and awareness of one another's mobile privacy and security behaviors, along with individual and community participation in mobile privacy and security co-management, increased from pre- to post-study. Interview findings confirmed that the app features supported collective considerations of apps and permissions. However, participants expressed a range of concerns regarding having community members with different levels of technical expertise and knowledge regarding mobile privacy and security that can impact motivation to participate and perform oversight. Our study demonstrates the potential and challenges of community oversight mechanisms to support communities to co-manage mobile privacy and security.

1 Introduction

The majority of U.S. adults own smartphones [50], and nearly half of them have reported downloading various third-party apps [8]. These mobile apps often require access to users' sensitive information, such as contacts, emails, location, photos, calendars, and even browser history [8]. Most apps request users' permission before accessing any information or resources. Yet users may have difficulty understanding these permission requests and the implications of granting them [7, 29, 51]. As a result, users struggle to make permission decisions or grant permission by mistake [35]. Even worse, there are ways for more malicious apps to circumvent the permissions system and secretly gather users' system resources and private information without consent [56]. Ironically, a recent Pew Research study reported that most US adults are concerned about how their personal information is being used by these third-party apps as respondents felt they lack control over their mobile privacy [22, 63].

This lack of understanding leads users to seek advice and guidance from others [24]. Several studies have demonstrated that users often learn about privacy and security from their social network, which influences them to change their own digital privacy and security behavior [27, 46, 60]. As such, networked privacy researchers acknowledged the importance of these social processes for managing individual and collective digital privacy and security [20, 45, 54]. Despite this prior work, few mechanisms to support these social processes have been developed and evaluated. In this paper, we explore community oversight, where trusted groups of users help one another manage mobile privacy and security. In our previous work, we proposed a theoretical framework of community oversight [17], describing how the concepts of transparency, awareness, trust, individual and community participation are needed within a particular mechanism. We have now implemented a mobile app, Community Oversight of Privacy and Security (CO-oPS), to explore these concepts in use and support a collaborative approach to mobile privacy and security management. The CO-oPS app allows individuals in a community to review one another's apps installed and permissions granted and provide direct feedback to one another.

In this paper, we present a field study of the CO-oPS app. Our aim was to understand the impact of using the app on participants' mobile app decisions and perceptions. We conducted a 4-week mixed-method longitudinal field study with

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.
USENIX Symposium on Usable Privacy and Security (SOUPS) 2023.
August 6–8, 2023, Anaheim, CA, United States.

101 people in 22 self-formed groups. Each group installed, used, and evaluated the CO-oPS app, provided oversight to one another on their mobile app privacy decisions, and shared experiences through weekly surveys and optional interviews. We describe how users interacted within the app and the changes in their mobile app permission decisions after using the CO-oPS app. We also examine how participants' perceptions regarding co-managing their mobile privacy and security within their communities change throughout the study. To do so, we measured constructs derived from our community oversight model [17] of perceptions of transparency, awareness, trust, and individual and community participation within the CO-oPS app. We tested for the pre-post study differences and detected increases for all of these measures that were statistically significant. Qualitative findings further explain these perceptions and identify co-management concerns: feelings of privacy invasion of their own and others, lack of trust in less knowledgeable community members, lack of close relationships, and communities' inadequate tech expertise. We also found that using the CO-oPS app helped participants increase their communities' collective capacity to address their mobile privacy and security concerns.

In sum, our study makes a unique contribution to SOUPS research community by investigating through a field study how a community oversight mechanism can help increase participants' collective capacity to support one another in co-managing mobile privacy and security together as a community. Specifically, we make the following unique research contributions: 1) Through a longitudinal field study, we describe the benefits and challenges of using a community oversight app to co-manage mobile privacy and security; 2) We provide empirical evidence of the potential for community oversight to increase users' awareness of mobile privacy issues, leading to individual changes in decisions and community exchange of knowledge; and 3) We present considerations and design-based recommendations towards features to support communities in providing oversight to one another.

2 Background

Privacy and Security Management in Mobile Applications

Mobile applications often access sensitive information and share users' personal data with third parties [13,26,32,42,56]. As such, substantial work has been done to investigate and support end users in managing mobile app privacy and security. Researchers have looked at the existing privacy awareness and management approaches (e.g., app privacy permission prompts, privacy policies, etc.) and found that such mechanisms often fail to provide users with awareness and knowledge of privacy and security risks [6, 27, 29, 35, 62]. Moreover, users often do not understand mobile app permission dialogues [29] and are over-exposed to such requests [62]. Researchers have proposed several technology-based solutions to increase awareness and limit potential risks associated with third-party mobile apps [43, 52, 58]. For example, Sadeghi et al. suggested evaluating the app permissions against risks and automatically grant/revoke permission on users' behalf [58]. Others proposed mechanisms to inform users about the app privacy risks, recommend secure choices, and nudge them to review/revise permissions [6, 40, 67]. Others suggested tools to allow users to review data before sending it to the server, visualize data flow [9], and replace personal information with mock data without affecting app functionality [43].

While this body of research has emphasized enhancements to technology to help individuals manage privacy and security while using mobile applications, none looked at how knowledge and influence from social groups help in individual privacy and security decision-making. Our research focuses on assessing and supporting these social processes involved in privacy and security management.

Community-based Approaches for Privacy and Security

In general, research shows that people frequently take collaborative approaches to make privacy and security decisions [49, 54], and users often rely on social factors while making such decisions. Chin et al. discovered that smartphone users are more likely to consider social signals, such as reviews and ratings from other users, rather than privacy indicators regarding Android permissions when making app use decisions [16]. Das et al. demonstrated that social factors (e.g., community adoption of security features) could increase individuals' security awareness and encourage them to adopt security features [21]. As such, researchers have proposed using social and community influence to assist individuals in making decisions about digital privacy and security [31, 45, 61]. Squicciarini et al. developed CoPE, a tool to support users in collaboratively managing their shared images in social network sites [61].

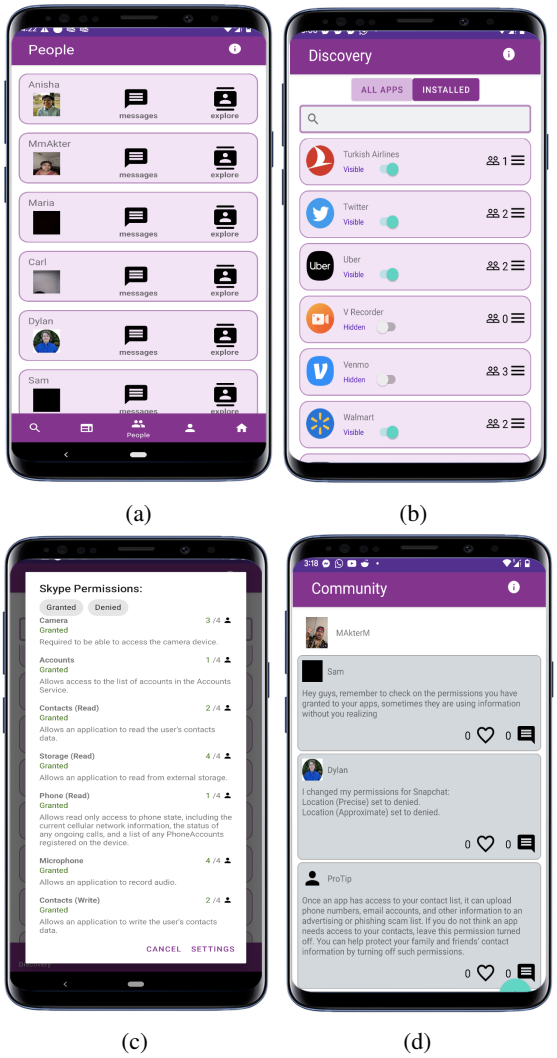
Past research has also examined privacy management approaches involving one party performing oversight for another. Organizations adopt mobile device management (MDMs) systems to remotely control and secure the data stored in employees' mobile devices [33]. Parents use adolescent online safety apps to monitor and protect teens by restricting their online behavior [1,4,30,65]. The results from these studies suggest that a collaborative approach, rather than one-sided control, could benefit both parties and lead to more privacy-preserving outcomes. Finally, several studies leveraged crowdsourcing to use mass user data to support individual users in making improved mobile privacy and security decisions [34, 38, 41, 55, 66]. For instance, Ismail et al. utilized crowdsourcing to recommend permissions that can be disabled for enhanced privacy without sacrificing usability [34]. However, these approaches showed little consideration for the trustworthiness of information from a random crowd. On the other hand, researchers found that users are more willing to adopt and share privacy advice from a trusted community [57], and they often communicate first with friends and family to learn about potential privacy and

security threats and mitigation strategies [20].

In summary, our work builds upon the past literature in social cybersecurity, MDMs, parental control apps, and crowdsourcing to implement and evaluate a novel model of community-based oversight (i.e., self-selected groups) for mobile privacy and security through a large-scale field study. Since the network structure of oversight (e.g., individual for MDMs, many-to-one for crowdsourced recommendations, and unidirectional from parent to child for parental control) in these prior works is vastly different than ours, this new model of community oversight warrants deeper empirical investigation. In [17], we were the first to propose a novel framework of community oversight for helping people manage their mobile privacy and security together. Through a participatory design study, we identified mechanisms that would allow users to support others in the community in making privacy and security decisions regarding mobile app permissions. We also designed a prototype mobile app that allows users to collaborate and share information with people they know to help make mobile app permissions decisions [5]. While this body of our prior studies provides a valuable basis for the design of community-oriented privacy and security management systems, they only present a theoretical view of users' preferences in community decision-making. In contrast, this study contributes to the literature by providing an in-situ evaluation of how trusted groups of people use and interact with different community-oriented features to collaboratively manage their mobile privacy and security.

3 Design of The CO-oPS App

We developed the Community Oversight of Privacy and Security (CO-oPS) Android app [2] based on the model of community oversight proposed in our prior work [17]. This model outlines the need for community oversight mechanisms to support individual and community participation through awareness and transparency features that build trust between community members. Thus, our CO-oPS app design includes four key features: 1) People page, 2) Discovery, 3) Permissions, and 4) Community Feed. The Discovery page allows community members to review one another's installed apps (Figure-1(b)), and the list of permissions granted or denied to each app (Figure-1(c)). Users also can review the count of total community members who have the same apps installed or permission granted. To help users change the app permissions easily, the Permission page provides a "SETTINGS" link that forwards users to Android Settings to modify app permissions. On the Discovery page, users can also hide some of their own apps from their community, ensuring their personal privacy. To provide feedback to one another, users can direct message and can openly discuss any privacy and security issues on the Community feed page (Figure-1(d)). This community feed has another important function: when someone in the community changes their app permission, the CO-oPS app creates



(a) (b) (c) (d)

Figure 1: CO-oPS Features: (a) People, (b) Discovery, (c) Permissions, (d) Community Feed.

an automatic post on the community feed about that change. It also posts weekly protips to educate community members regarding safe apps and permissions.

4 Study Constructs

To evaluate the impact of using the CO-oPS app, we measured a set of constructs that we surveyed before, during, and at the end of the field study. We measured all constructs by presenting participants with various statements relevant to each construct. Participants were asked to rate each statement on a 5-point Likert scale from 1 (strongly disagree) to 5 (strongly agree). First, we developed new constructs derived from the theoretical framework for community oversight proposed in our prior work [17], consisting of transparency, awareness,

trust, individual participation, community participation, and community trust. We validated these new constructs through standard psychometric tests (i.e., Cronbach's alpha [19] to confirm internal consistency), which is reported in Table-5. Then, we utilized three pre-validated scales from prior research [14, 15, 36, 59] to measure community belonging, self-efficacy, and community collective efficacy. All scale items are included in Appendix A. Below, we define each of the constructs, along with our hypotheses.

Transparency: As Das et al. demonstrated [21], social proof - seeing others adopt a privacy and security behavior - often helps individuals adopt the same behavior. Therefore, to encourage individuals in a community to make informed decisions for their mobile privacy settings, the behaviors of others must first be transparent. Therefore we define transparency as an individual's perceived visibility of their community's mobile apps installed and the permissions granted/denied.

H1: *At the end of the study, community members will perceive higher levels of transparency in their community's mobile privacy and security behaviors.*

Awareness: Endsley demonstrated [25] that situational awareness - the understanding of what is going on around someone - is a key component in effective decision-making. In a later study [23], DiGioia and Dourish suggested that being informed about digital privacy and security norms and practices along with the actions performed by the community are necessary for an effective social influence process. We developed our awareness measure as an individual's perception about the awareness of their own and others' apps installed, permissions granted/denied, along with the changes made.

H2: *At the end of the study, community members will perceive higher levels of awareness regarding their community's mobile privacy and security practices.*

Trust: In [17], we identified that having the information available and being informed about mobile privacy and security practices might not be sufficient for community oversight. This is because individuals need to be able to trust the quality of the information and perceive the information as dependable to learn from and be influenced by it.

H3: *Community members will have a higher level of trust in one another's mobile privacy and security decisions.*

Individual Participation: While an effective social process needs transparency, awareness, and trust in one another, individuals also need to be willing to engage in this process [17]. Users need to be motivated to utilize the knowledge gathered from their community in order to make decisions. They also need to be willing to provide oversight to others. Thus we define individual participation as an individual's willingness to take steps to make changes in their own mobile privacy and security behaviors (uninstalling unsafe apps or denying dangerous permissions) and also providing oversight to others' mobile privacy and security behaviors (providing feedback and guidance to others).

H4: *Community members will perceive higher individual participation at the end of the study.*

Community Participation: Community oversight mechanisms can take place in different types of communities, such as, families [18], coworkers [39], friends, and social networks [39]. Yet not all types of communities may have an equal level of willingness to take part in different forms of community oversight. For example, in [17], we found that communities with closer relationships might be more willing to help one another make decisions than communities with weaker ties. Therefore, we define community participation as an individual's perception of their community to collectively work together, e.g., help one another, exchange feedback and guidance, and engage in open discussions.

H5: *At the end of the study, participants will perceive a higher level of community participation.*

Community Trust and Belonging: Individuals are likely to help one another if they feel like they belong and can trust their community members. We define *Community Trust* as an individual's perception of trusting their community to keep their personal information (e.g., apps installed) private and care for one another's mobile privacy and security. For community belonging, we utilized a pre-validated measure [14, 59] that has been used in exploring community support mechanisms outside of privacy and security. The *community belonging* construct measures an individuals' feelings about how much they matter to their community. While our participants already knew each other, participating together in the CO-oPS app could lead them to feel stronger bonds and care between each other. Therefore, our hypotheses are:

H6: *An individual's community trust will be higher at the end of the study.*

H7: *Community belonging will be higher after the study.*

Efficacy: Two of the outcomes we wanted to measure are perceptions over the efficacy of individuals and groups to manage their mobile privacy and security. Thus, we used pre-validated measures for self-efficacy [10], and community collective efficacy [15] in our study. The *self-efficacy* [10] construct measures an individual's perceived capacity to manage their own mobile privacy and security. The *community collective efficacy* [15, 36] construct measures an individual's perceived collective capacity to manage their community's privacy and security together. Our hypotheses are:

H8: *Individual's self-efficacy will be higher after the study.*

H9: *Community collective efficacy will also be higher at the end of the study.*

5 Methods

Study Overview: The overall goal of our study is to evaluate the CO-oPS app in building the capacity of the communities to manage their mobile privacy and security collectively. We also wanted to understand what impacts this community-based approach may have in changing participants' perceptions and behaviors toward their individual and collective mobile privacy and security management. To achieve these goals, we recruited small self-organized communities (2-6 Android phone users) who knew each other. Each community member installed the CO-oPS app and participated for four weeks. Measures were gathered before app installation, each week of the study, and at the end. Each week participants were asked to complete different in-app tasks that allowed them to explore the features of the CO-oPS app. Finally, participants were invited to participate in an optional follow-up interview. In each step of the study, we explicitly provided the definition of the term "community" as "your group members who are participating in this study." Each participant was compensated with a $40 Amazon gift card for completing the field study, with an additional $10 Amazon gift card for participating in the interview. Some participants withdrew from the study after two weeks due to technical difficulties with their smartphones and were compensated half the amount. Twenty-nine participants discontinued participation after week one, perhaps due to natural attrition, and were not compensated. Data were discarded from all who did not complete the study.

Table 1: Sociodemographic Characteristics of Participants

	Total no. of participants	*N=101*	*100%*
Gender	Female	46	45.5
	Male	55	54.5
Age	13-17	6	5.9
	18-24	27	26.7
	25-34	49	48.5
	35-44	6	5.9
	45-54	10	9.9
	55-64	1	1
	65+	2	2
Ethnicity	Asian/Pacific Islander	72	71.3
	Black/African American	13	12.8
	Hispanic/Latino	8	7.9
	White/Caucasian	8	7.9
Education	Primary School	8	7.9
	High School	5	5
	College (Associate)	6	5.9
	College (Bachelor)	40	39.6
	Masters	36	35.6
	Doctorate	6	5.9

Participant Recruitment: We recruited a total of 101 participants that were associated with 22 communities. We initially recruited the primary contacts of each community who completed a pre-screening eligibility survey that verified whether they met the inclusion criteria of the study prior to providing their informed consent. The inclusion criteria for participation included: 1) reside in the United States, 2) be 13 years or older, 3) have an Android smartphone, and 4) be willing to install and use the CO-oPS app. Here, we also specified that they "must participate in a group with two other people you know," which determined the minimum group size required to participate in this study. After completing the screening survey, the initial contacts were asked to share this eligibility survey with people they knew to invite them to participate in this study as their community members. Therefore, the initial contact of each group self-selected their community based on the above criteria (1-4). As such, all group members knew the initial contact but in some cases, were only loosely acquainted with one another. For the teen participants, we required their parents to complete this survey and provide their consent.

Our study was Institutional Review Board approved. The target characteristics of our participants were all Android smartphone users of any age range (minors, adults, and older adults). Therefore, we did widespread recruitment through social media, email, phone calls, and word-of-mouth. The recruitment process started in January 2022 and ended in August 2022. Overall, we recruited 22 communities (101 participants) where the size of the communities ranged from 2 to 6. Table-1 summarizes the gender, age groups, ethnicity, and education of our participants. Our participants were primarily young, between the ages of 13 to 34. Most of them had a college degree. The majority of the participants were Asian, followed by African American, Hispanic/Latino, and White/Caucasian. Table-2 illustrates the frequency of the group compositions. Most of the groups consisted of family members, friends, and others (e.g., neighbors, co-workers, and acquaintances).

Table 2: Group Compositions

Total no. of groups	*22*	*100%*
Family Only	2	9.1
Family and Friends	4	18.2
Family, Friends and Others	8	36.4
Friends Only	3	13.6
Friends and Others	4	18.3

App Tasks: During the field study, our participants were asked to explore different parts of the CO-oPS app through a set of tasks each week. These tasks prompted them to become familiar with CO-oPS features and introduced them to the goal of collaboratively managing mobile privacy and security. Table-3 depicts the weekly tasks. For example, Week 1 tasks asked participants to become aware of their own mobile privacy and security decisions, whereas Week 2 tasks asked them to perform oversight of others in their community. Participants could check off completed tasks in the app to remove them from their task list, but we otherwise did not track or require completion to continue in the study.

Table 3: Weekly App Tasks

Week-1: 1) Review your own apps from the "Discovery" page > "Installed" tab. Hide the apps that you do not want others to see. 2) Review your community's apps from the Discovery > "All Apps". Check if you have any uncommon apps that no one else is using. 3) For the apps you have in common with others, compare the permissions you granted but others denied.
Week-2: 1) Read the weekly pro tip and add a comment there. 2) From the "People" page, review one of your community member's apps. Check if there are any apps or permissions that may not be safe. 3) Send a message to warn them about unsafe apps or permissions.
Week-3: 1) Read the weekly pro tip and add a comment there. 2) Review your own apps and permissions and check if any granted permissions may be unsafe. Consider changing those permissions. 3) Review the apps and permissions of someone in your community. Let them know if they have any unsafe permissions.
Week-4: 1) Check the messages received from your community. Consider changing the apps and permissions accordingly. 2) Review your community members' apps and check if any unsafe apps or permissions exist. 3) Write a post on the Community feed to warn others about the unsafe apps or permissions found.

Survey Design: Each participant completed two Qualtrics surveys (pre-study and post-study) before and after the field study, which contained four constructs: self-efficacy, community belonging, community trust, and community-collective efficacy. The pre-study survey also collected participants' demographic information, e.g., age, gender, ethnicity, and education. During the field study, participants also completed a shorter Qualtrics survey each week (weekly survey), containing all constructs of the community oversight model. Links to the weekly surveys were delivered through the CO-oPS app, which redirected participants to the Qualtrics web survey.

Follow-up Interview: At the end of the field study, we invited participants to an optional 30-minute one-on-one interview session on Zoom to learn about their experience using the CO-oPS app with their community. Fifty-one participants from 18 communities participated in the follow-up interviews. We started the semi-structured interview by asking about mobile privacy and security practices before participating in the study. Next, we asked about their overall experience of using the CO-oPS app. Participants were also encouraged to express their perceived benefits and concerns about different features of the CO-oPS app. Appendix B presents some sample interview questions we asked during the follow-up interviews. The interview sessions ranged from 40-70 minutes and were audio/video recorded.

Data Collection and Analysis: The study produced a rich dataset: 1) quantitative data from survey measures, 2) CO-oPS app usage logs, and 3) qualitative data from follow-up interviews. We first categorized the survey responses as pre-study, week-1, week-2, week-3, week-4, or post-study, depending on the timestamps of the survey completion. Then, we verified the construct validity of our measures using Cronbach's alpha [19] and created sum scores to represent each construct. Next, we conducted Shapiro–Wilk tests and found that the sum scores of the constructs were not normally distributed ($ps < .01$). Therefore, we performed the non-parametric Wilcoxon rank-sum test to identify significant differences between the pre-study and post-study measures (Table-5). We also present the descriptive statistics for each pre- and post-study survey item of the newly developed constructs (Appendix D)

We instrumented the CO-oPS app to log participants' usage data. We also stored the list of the apps installed and permissions granted/denied during the installation of the CO-oPS app and at the end of the field study. We analyzed the usage log to identify how and at what frequencies participants utilized different features of the CO-oPS app. We also analyzed the pre- and post-study app/permissions lists to investigate the changes made to the apps and permissions during the study. Due to some technical issues with the CO-oPS logging feature, we could not log the in-app activities of the first seven communities. Therefore, the app usage data was received from only the last fifteen communities (N = 68 participants).

We qualitatively coded our interview data using inductive analysis techniques [28] to understand how participants perceived the CO-oPS features that tie to the constructs we were measuring. Thus, our qualitative analysis complemented the quantitative results from our surveys. We first familiarized ourselves with our data by reading through each transcript and then template-coded our data based on the community oversight concepts described in Section 4. Specifically, we coded for 1) the level of transparency on the information shared by them or others, 2) the types of information they felt helped raise their awareness about the community's mobile privacy and security practices, 3) the level of trust of one another's privacy behaviors and advice, 4) how and whether they would individually participate in such a community, 5) how participants discerned community participation, 6) the trust and belonging they felt with their communities, and 7) their individual and community-level capacity to manage mobile privacy and security. The first author worked closely with three researchers to code the data iteratively and formed a consensus among their codes. The remaining authors helped guide their analyses and interpretation of the results. Appendix C presents the codes and illustrative quotations for each analysis theme.

6 Results

On average, participants spent 32 minutes in the CO-oPS app over four weeks, ranging from 19 minutes to 1 hr 31 minutes. Table-4 summarizes the activity types that participants performed with the CO-oPS app. Table-5 summarizes Cronbach's alpha, means, standard deviations, skewness, and kurtosis of each construct measured during the study. All

Table 4: CO-oPS App Activities

Activity Types	*N = 68*	*100%*	Average*	Apps/Permissions Types
Hide installed apps	34	49%	6	Games, video streaming, banking, online shopping
Review own app permissions	61	87%	23	Games, online shopping, social media, mobile payments
Review others' app permissions	46	65%	18	Social media, banking, games
Send private messages	51	74%	1	Apps: games, social media; Permissions: locations, camera, microphones, contacts

* Average number of activities per user (on apps, permissions, or messages)

Cronbach's alphas were greater than 0.80, which suggests good internal consistency of our measures. Next, we tested for within-group differences in the constructs based on whether the participants completed it at the start or end of the study. We will discuss each measure below, along with the corresponding findings from the qualitative data and usage logs related to each construct.

Transparency: As shown in Table 5, participants reported higher ($p = .010$, M1=4.07, M2=4.36) levels of transparency (an individual's perception of whether CO-oPS gave them a transparent view of the apps installed and permissions granted on their community's mobile devices) at the end of the study. Hence, this result supported our hypothesis (H1). Our qualitative results also confirmed that almost all participants felt that CO-oPS made their community's mobile privacy and security decisions visible to them. Three-quarters of the participants interviewed (76%, N=39) explicitly said they liked the CO-oPS feature that let them **check own apps and permissions**. Participants often mentioned that having the ability to see all their installed apps on one screen provided them transparency of their app usage. Two-thirds of the participants (67%, N=34) also brought up the **visibility of others' apps and permission**. To this end, they said reviewing others' apps and permissions provided them a sense of purpose for using CO-oPS with their communities. Interestingly, while these participants appreciated the ability to review one another's apps and permissions, they often referred to the importance of the CO-oPS app-hiding feature because it made them feel less intrusive to others' privacy. As such, C18P1 said, *"Because some of the apps can be hidden if someone likes, that gives me the feeling of a relief when I see others' [apps installed]."*. However, some participants (25%, N=13) believed that this app-hiding feature **defeats the main purpose** of CO-oPS.

Some participants, on the other hand, perceived transparency as a two-way privacy violation, e.g., the privacy of themselves and the privacy of others. For example, more than one-third of the participants (38%, N=20) felt that **their personal privacy was being violated** as others, who were not close, could see their personal information (e.g., installed apps and permissions). Some participants (27%, N=14) also specifically said **they might forget to hide an app after installation**, which could leave their apps visible to others. On the contrary, one-fourth of the participants (25%, N=13) felt that this transparency of others' apps and permissions **can be privacy-invasive to others** as well. C11P2 said, *"While using the app, my friends and I discussed privacy more than security because we can see the apps on our friend phones and I think that's not a good thing."*.

The results from the log analysis (Table-4) also supported the above concerns. For instance, around half of the participants (49%, N=34) hid one or more of their installed apps from their communities. Participants, on average, hid six mobile apps ranging from one to 17. The most frequent types of apps that participants hid were games, video streaming apps, banking apps, and online shopping.

Awareness: Participants overall reported a higher ($p < .001$, M1=3.95, M2=4.37) level of awareness (individual's perception of whether CO-oPS made them more aware of their community's mobile privacy and security decisions) after the study. Hence, this result supports our hypothesis (H2). Our qualitative results showed that using the CO-oPS app helped participants raise their overall awareness of mobile privacy and security issues, along with their awareness of one another's privacy and security practices. For example, almost all participants (94%, N=48) felt that they became more **aware of mobile privacy and security issues** because CO-oPS enabled them to focus on permissions. They also became more aware of which of their personal information was being accessed by their installed apps. For example, C15P1 said, *"It just makes it more obvious. It's very focused on permissions. So I think having that focus, it's very beneficial. People in the community, I see are now more concerned... for their permissions specifically. I totally see how it changed our perspectives."*. Some of these participants (39%, N=20) often brought up the **weekly pro tips** they got on the CO-oPS app as it helped them increase their awareness regarding mobile privacy and security in general.

Most participants (86%, N=44) also said they became more aware of whether **a permission is necessary for an app**. They often mentioned that comparing their own app permissions with others helped them increase this knowledge. Almost all of these participants (82%, N=42) also mentioned that it helped them **keep track of their own apps**, as they found more installed apps on CO-oPS' Discovery page than they were aware of. They also often mentioned some granted permissions found on the CO-oPS app (e.g., microphone, camera, location, contacts, etc.) that they did not remember granting.

Around half of the participants (57%, N=29) said they **be-**

Table 5: Descriptive Statistics and Wilcoxon Rank-Sum Tests of Pre and Post-Study Responses to the Constructs

Constructs		Pre-Study				Post-Study					
	α	*M1*	*SD1*	*S1*	*K1*	*M2*	*SD2*	*S2*	*K2*	*V-val*	*p-val*
Community Oversight Model:											
Transparency	0.88	4.07	0.80	-0.74	0.70	4.36	0.67	-0.74	-0.49	1054*	**0.010**
Awareness	0.82	3.95	0.79	-0.76	0.84	4.37	0.68	-0.72	-0.45	1110.5***	**<0.001**
Trust	0.90	3.61	0.80	-0.29	0.11	4.28	0.85	-0.81	0.11	633.5***	**<0.001**
Individual Participation	0.87	3.78	0.83	-0.67	0.59	4.23	0.73	-0.89	0.17	897.5***	**<0.001**
Community Participation	0.88	3.86	0.80	-0.56	0.24	4.18	0.83	-0.92	0.23	1002**	**0.002**
Community Trust	0.87	4.03	0.87	-0.85	0.45	4.22	0.71	-0.78	-0.03	1412*	**0.048**
Community Belonging	0.91	4.09	0.67	-0.53	-0.52	4.20	0.68	-0.60	-0.88	1899	0.209
Community Collective Efficacy	0.91	3.80	0.78	-1.09	2.56	4.12	0.68	-0.44	-0.36	1287***	**<0.001**
Self Efficacy	0.85	3.95	0.64	-0.58	0.79	4.32	0.61	-0.80	0.27	1021***	**<0.001**

*p<.05; **p<.01; ***p <.001

came more aware of their community members' privacy and security behaviors. Here, they mostly mentioned one or two people in their community whose apps they could keep an eye on to ensure their safety. Lastly, some participants (39%, N=20) also said that they appreciated the CO-oPS feature that **informed everyone about the permission changes** made by any member as this helped them decide whether to imitate that change. Around one-fifth of the participants (18%, N=9) felt that CO-oPS app **did not make them aware of the app changes** made in the community - the apps community members installed or uninstalled on their phones. This was not a feature we implemented, and these participants felt like it had been overlooked and desired that awareness.

The findings from our log analysis (Table-4) are reflective of the quantitative results. For example, most of our participants (87%, N=61) checked their own app permissions during the study. Participants, on average, reviewed the permissions of 23 apps and primarily explored the permissions of apps that are about gaming, online shopping, social media, and financial payments. Alongside reviewing their own apps and permissions, more than two-thirds of the participants (65%, N=46) reviewed others' app permissions. On average, they explored 18 app permissions of their community members. The most common types of apps being reviewed were social media, banking, and gaming apps.

Trust: Similar to the above two constructs, the post-study responses saw a higher ($p < .001$, M1=3.61, M2=4.28) level of trust (individual's perception of whether CO-oPS helped them foster trust in one another's mobile privacy and security decisions) among the community members. This result confirmed our hypothesis (H3). Almost half of the participants (51%, N=26) said they **found the advice provided by their community was dependable**. They were overall appreciative of the feedback and guidance they received from more tech-savvy community members, as it helped them learn more about risky apps and unnecessary or dangerous permissions. However, the trust did not always extend to all community members. For example, C18P1 said: *"In this app, you're trusting each other's decisions. But for me, in this community, only [Name] is more tech-savvy. And most of the people are not. And these decisions are not always well-informed, right? So, I follow only [Name] to check what he has."*.

Conversely, participants felt that trusting others' privacy and security practices might be challenging in some cases. Around half of the participants (49%, N=25) said **some of their community members were less knowledgeable** about mobile privacy and security issues. Therefore, they could not trust those people's mobile privacy and security decisions to learn from. As C11P3 said, *"I don't think they were much of aware. They do not care of all this, you know, privacy and security stuff,... so I am not sure I followed them, their permissions and stuff."*. Interestingly, they also often mentioned that those with less knowledge were **less tech-savvy** (37%, N=19) in general.

Individual Participation: Participants reported a higher ($p < .001$, M1=3.78, M2=4.23) level of individual participation (perception of whether the CO-oPS app helped individuals participate in their own and others' mobile privacy and security decisions) after using the CO-oPS app for four weeks. This supported our hypothesis (H4). Our qualitative results also revealed that participants overall took the initiative to change their apps and permissions and also provided their oversight to others. Notably, more than two-thirds of the participants (67%, N=34) said that they **made changes to their own apps and permissions**. Participants often said that they made these changes after reviewing their own permissions and identifying the unnecessary or concerning ones by themselves. Some other participants said that comparing their own app permissions with their community's inspired them to change their app permissions. Some of these changes were made because of feedback received from other community members. C02P1 said, *"I did some changes. I denied some of my permissions. [Name] asked me to remove the microphone from one of the apps I use for workouts. I have removed it now. ... also, you can always just check and then you just have to learn what permissions are suspicious and what are necessary."*.

Next, more than one-third of the participants (41%, N=21) explicitly said that they **provided feedback to their commu-**

nity members to warn about the apps that they thought might be risky or the permissions granted that might be a cause of privacy concerns. To provide feedback, participants did not just use the CO-oPS messaging feature, they also mentioned using other media, e.g., text messages, social media private messages, phone calls, or talking in person.

Log results (Table-4) demonstrate that individuals did provide oversight during the study. We found that 74%, N=51 participants sent messages to someone in their communities, where twelve messages were about warnings regarding risky apps (games, social media). Thirty-five messages contained warnings regarding specific app permissions they found on their community members' phones. They mostly provided feedback about location, camera, microphones, and contact permissions. For instance, C09P1 messaged C09P3: *"You're granting Douyin a ton of permissions. maybe we should keep the Chinese spyware to a minimum."*

However, some participants expressed a number of factors that reduced their motivation to participate. More than one-third of the participants (41%, N=21) believed that they **were less tech-savvy than others in their community** and therefore they doubted their feedback would be useful to others. Interestingly, some participants (39%, N=20) felt that the **people who participated with them were not close** and therefore they did not care about those people's mobile privacy and security. A few participants (29%, N=15) expressed that they **had very few mobile apps installed** on their devices, and so, they did not need to be concerned about mobile privacy and security. Ironically, some of these participants also believed that they did not have anything to be concerned about because the personal information that is stored in their mobile phones is not very sensitive in nature. A few also felt that their information was already leaked by some online entities and so it was too late to start caring about mobile privacy and security. As such, C14P2 said: *"I don't see the point now because you can't just control what they [apps] already stole from you. I use very few apps, and all my data is already out there*."

Community Participation: The community participation measure (individual's perception of whether the CO-oPS app enabled the community to help one another make their mobile privacy and security decisions) increased ($p = .003$, M1=3.86, M2=4.18) over the duration of the field study. This confirmed our hypothesis (H5). More than three-fourths of the participants (78%, N=40) said the CO-oPS app **allowed them to learn from their community** regarding mobile privacy and security management and exchange their knowledge regarding app safety and privacy. Most of these participants (71%, N=36) also mentioned that using CO-oPS helped them **initiate more open discussions** regarding mobile privacy and security in their community than ever before. They said these discussions most often took place offline when they saw one another in different social gatherings. Around half of the participants (53%, N=27) specifically discussed **receiving feedback and advice from their community**. C17P1 said, *"I mean, offline, or virtually, we kind of worked together, we talked, we get each other's knowledge. But that also happened with the co-ops app, that there were so many options to get in touch with each other by that messaging or, notifying them, or community discussion... I will say it kind of, we helped one another learn as a team."*.

However, some participants said the CO-oPS app might not help increase community participation when the members are extremely or not particularly tech-savvy. One-third of the participants (31%, N=16) envisioned that **when the community members are less tech savvy**, they might not be able to provide oversight to each other. On the other hand, 27% of the participants (N=14) said that their **entire community was very tech-savvy and well aware of the mobile privacy and security issues**, and therefore they did not find it necessary to engage in discussion or exchange feedback with one another. C11P5 said: *"My community is from a computer science background. I think we are already aware of these things. So, we don't need others' advice."*.

Community Trust and Belonging: While community trust increased over the course of the study ($p = .048$, M1=4.03, M2=4.22), the difference between community belonging was not statistically significant ($p = .209$, M1=4.09, M2=4.20). Thus, hypothesis (H6) is supported, but (H7) is not supported. In our qualitative analysis, we found that all of our participants (100%, N=51) said they **personally knew each member** of their communities. Most of our participants (86%, N=44) mentioned **having close relationships**, e.g., family members, friends, co-workers, and neighbors, with some members of their communities. Thus, using CO-oPS did not appear to bring groups closer together.

However, perceptions of trust and community relationships were still important in how individuals interacted with each other in CO-oPS. Around half of the participants (47%, N=24) said that they **had trust in their community that their apps and permission information would not be misused**. One-fourth of our participants (24%, N=12) said they had peace of mind because they would **rely on their community members** who would actively monitor their mobile privacy decisions and warn them if anything is found concerning. Here, we often noticed that participants referred to some specific community members, not the entire community, who they would rely on. C02P1 said, *"With [Name] in my group, at least I know that if he saw something he didn't think wasn't proper, he will definitely let me and my husband know...We have that kind of relationship, so we know we can trust him."*.

However, a few participants felt that sharing the apps installed might cause some security issues due to the lack of trust in certain community members. For example, a few participants (18%, N=9) envisioned **security concerns in sharing their financial apps**, such as banking or mobile

payments, with their community. They often brought up hypothetical scenarios of a family member (e.g., children) knowing what apps they have installed, who would somehow get access to their phone, log in to their financial apps, and transfer money. A couple of participants also imagined situations when community members **might judge or bully them because of their choice of gaming apps**.

Self-efficacy: Our participants reported higher levels of self-efficacy (individual's capacity to manage their mobile privacy and security) at the end of the study ($p < .001$, M1=3.95, M2=4.32). This confirmed our hypothesis (H8). Most participants (80%, N=41) said they **gained confidence in managing their mobile privacy and security**, particularly by reviewing their installed apps and granted permissions and identifying whether there is anything concerning. C10P1 said, *"So, I can now think through it, like what is the purpose of this permission? Like if the permission conflicts with the purpose of the application, I can just turn it off. You see, this is new. I now can differentiate what's necessary or what's not."* Interestingly, more than half of the participants (57%, N=29) said they now have **become more knowledgeable about changing permissions**, mostly because they could easily navigate to the app permission settings from the CO-oPS apps. This perception was not universal, though. Around one-third of the participants (31%, N=16) also said they **already had the ability to manage their own apps and permissions** prior to participating in this study, and they never reached out to others for help.

Commmunity Collective Efficacy: Participants reported higher community collective efficacy (individuals' belief that their community can co-manage mobile privacy and security) at the end of the field study ($p < .01$, M1=3.80, M2=4.12). This confirmed our hypothesis (H9). Reflecting this, most participants (88%, N=45) felt they could easily **reach out to their community and work together as a team** for their mobile privacy and security decisions. Most of these participants (67%, N=34) mentioned that they **have at least one person in the community they could reach out to ask questions** about whether an app was safe to use or a permission should be allowed. C03P5 said, *"When I'm giving permissions, I now can tell that could be the things that are needed for a discussion. I do go to [Name] to ask what he thinks. what he thinks the permission is needed or not needed for the app. I do my permissions like this now."*

Behavioral Impact:
Our log analysis results provide further insights into participants' overall behavioral changes regarding mobile privacy and security. We found that 87% of the participants (N=61) changed at least one of their app permissions during the study. Participants, on average, changed 29 permissions, where all permissions were changed to "deny." They mostly turned off the permissions accessing their location (approximate and precise), camera, storage, and contacts. For instance, C15P4 changed the Location (Approximate) permissions of Chase, Snapchat, and Gyve apps installed on his phone. However, participants did not show a similar decrease in the number of apps they had on their phones. Around 78%, N=53 participants installed new apps, whereas only a few participants (16%, N=11) uninstalled any apps. Participants, on average, installed two new apps, where the most common types of apps were mobile payment, banking, online shopping, social media, and games. On the other hand, the participants who uninstalled their apps mostly discarded gaming apps along with a few spiritual, fitness, and dictionary apps from their phones. Perhaps learning what apps others in their communities were using provided participants with ideas for additional apps they would be interested in.

7 Discussion

While our prior work conceptually proposed community oversight as a mechanism for supporting privacy and security management [17], this work is the first field study to empirically examine the real-world feasibility of implementing community oversight as a mechanism for co-managing mobile privacy and security among trusted groups. Our results largely confirm what was envisioned in that prior work: that community oversight does have the potential to help people help each other when it comes to decisions about mobile apps and app permissions [17]. Users' perceptions of their own and their community's capabilities to manage their mobile app privacy and security increased as a result of the study. The majority of participants modified their permissions, reducing what they were sharing with apps, and stated that their awareness of permissions and mobile apps also increased. Below we further discuss our overarching findings and their implications for the design of community oversight mechanisms.

Building Community Collective Efficacy
The goal of the CO-oPS app, as with many collaborative systems, is to build and support the collective capacity of groups to work together to achieve a common goal, in this case, to manage apps and app permissions. Thus, building community collective efficacy for mobile privacy and security is the primary end goal of CO-oPS. To that end, we believe our study was successful. The interview comments suggest that the community oversight mechanism helped our participants increase their ability to support each other in their mobile privacy and security decisions. Participants mentioned their change to a more collaborative perspective: the app facilitated knowledge sharing amongst their community and an ability to rely on others to help in decision-making. Our results also provide an empirical validation of the components of the community oversight model [17]. Again, both survey

and interview results demonstrated the roles of transparency, awareness, trust, and participation in providing community oversight. Future work could examine what factors are most related to community collective efficacy and thus are most important to provide in a community oversight mechanism.

Role of Tech Expertise

One of the key themes was that the level of tech expertise among community members plays a key role in bolstering or hindering community oversight. For instance, our participants expressed concerns about the potential lack of participation in communities when most members are sufficiently tech-savvy or knowledgeable about mobile privacy and security. Others expressed concerns about there being a lack of knowledge in their communities and less trust in the decisions of those with less expertise. Kropczynski et al. [36] also noted the importance of those with tech expertise in older adult communities for spreading privacy and security knowledge, even among those with low self-efficacy. This suggests that community oversight mechanisms may be most beneficial and appropriate when there are asymmetrical relationships among the community members in such a way that some community members need support while other members could provide that support. A key challenge is then how to incentivize those with sufficient expertise to participate in such communities, particularly to help community members they are not as close to or not already providing tech care to [37].

However, when this asymmetry in expertise combines with a power imbalance, which is often seen in families, the collaborative joint oversight might cause tension. Akter et al. [4] demonstrated that although teens had more expertise than their parents, they did not feel empowered to oversee their parents because of the existing power hierarchies. In families, parents often use parental control apps, a more restrictive approach that fosters monitoring and surveillance to ensure teens' mobile online safety, privacy, and security. Teens often perceive this unidirectional oversight mechanism as overly restrictive and privacy-invasive [30, 65]. Therefore, adolescent online safety researchers emphasize adopting a softer version than parental control or community oversight - a middle ground that allows parental oversight with bidirectional communication and teens' self-regulation [65]. So, the community oversight mechanism might need to incorporate additional features to help such unique types of communities with asymmetries in expertise and power.

Tensions around Transparency and Privacy

Another common concern was privacy issues arising from transparently sharing apps and permissions with others. While many appreciated such transparency, participants regularly chose to hide certain apps from other people. Some participants found this transparency too invasive and anticipated potential problems resulting from others knowing about what apps they use. Other concerns also arose from being able to determine if the advice given to another was taken or not, based on whether someone's permissions remained the same or changed. These concerns will likely be elevated as community size grows, where communities contain more members who are not close to one another. A recent study that explored collaborative mobile privacy management among families also found similar results where participants expressed concerns in including extended families with distant relationships [3]. To resolve these tensions, as with many collaborative systems, users may want more granular controls on who can see what apps and permissions rather than sharing equally throughout the community.

Incentives to Participate

Prior work identified that users might not be motivated to provide oversight to those not close to them [17] or those outside of existing care relationships such as between parents and teens [4]. Indeed, some of our participants expressed similar sentiments and were not concerned about the decisions made by those not close to them. Despite this, the majority of participants did perform oversight, and many interviewees described discussions and behaviors that were sparked as a result of that oversight. Yet with some incentives, participating in a user study, in this case, individuals performed the oversight, benefiting other community members. Thus a key question remains as to how to incentivize such oversight to different community members and how those incentives may need to change over time.

Implications for Design

Our results demonstrate how features that provide transparency and awareness and support trust between community members are essential components of community oversight. Mechanisms must also enable and encourage individual and community participation in the collaborative efforts of privacy and security management. Our results provide further insights into the features and mechanisms needed in a tool for communities to participate in collaborative oversight of their mobile privacy and security.

Making Privacy Features Visible: While the CO-oPS app had a feature that allowed users to hide any of their installed apps from others, it often failed to provide users with a sufficient sense of privacy. This may be because they were not well aware of this feature or were unsure how well it functioned. Participants also reported concern over forgetting to hide apps as they install new ones. Thus, mechanisms to keep users aware of this app-hiding feature will be necessary. Das et al. [21] and DiGioia and Dorish [23] also emphasized the importance of visibility so that users can be aware of the availability of the security feature and adopt it. To help users be aware of this feature, users can be prompted regularly or upon installing new apps to ask if they would like to hide. If community members hide too many apps, however, oversight

will be more limited. Thus, designers should also explore additional privacy features that can protect an individual's privacy while still allowing useful sharing to the community.

Raising Mobile Privacy Knowledge: One of our findings suggested that participants would not trust the mobile privacy and security behaviors of people who were less knowledgeable. This suggests that collaborative decision-making would not effectively function when there is little trust within the group. Increasing trust within communities may be very challenging, and how to do so remains an important open question. In [17], participants also envisioned such situations and recommended including external expert users whom the community members can turn to for guidance when they do not have the necessary expertise. Several other networked privacy researchers also demonstrated the need for knowledgeable expert stewards [11, 12, 48]. Therefore, we recommend app designers explore ways to include mobile privacy and security experts in communities. Another possibility is, rather than bringing experts into the community, to raise the expertise of certain motivated community members. This could include nudges towards additional information or resources, possibly personalized to those most amenable to such additional knowledge.

Increasing Community Participation: We found that our participants expressed several concerns about community motivations to provide oversight to one another. Individuals and communities, as a whole, need incentives to utilize a community oversight mechanism and continue to support each other [5, 53] in their knowledge-sharing and decision-making. Such needs for incentivizing individual participation in communities to support collective participation were also suggested by Watson et al. in [64] and Moju-Igbene et al. in [47]). Therefore, community oversight mechanisms need to include features that encourage such engagement and make the engagement of others apparent. For example, community members can be notified of any new apps installed or permissions granted on anyone's phone. Moreover, nudges could remind community members to review random members' apps and permissions. Additionally, lightweight feedback features might also help users to engage more. For instance, instead of messaging, users might prefer just to flag unsafe apps/permissions to notify others quickly.

Limitations and Future Work

We would like to highlight the limitations of our study that should be addressed in future work. First, our sample was skewed toward Asian adults, most of whom completed college and graduate-level education. Therefore, our results may not be generalizeable to other communities of different ethnicity, education, and age groups. Future work should explore communities with broader demographics, ethnicity, and socio-economic status [44]. Another limitation is that we asked our initial participants to form their communities with people they knew, which sometimes led to groups where everyone did not have strong bonds with each other. This may have led them to evaluate our app differently than if we studied with communities of families or close friends only. However, this also provided important insights into the importance of community trust in fostering oversight. Future work should examine how factors of group structure and relationships, including group size and varying levels of expertise, impact the motivation of participation and oversight activities of community members.

Although our qualitative results suggested that the CO-oPS app supported all necessary components of community oversight, this does not imply that our participants perceived usefulness, ease of use, and behavioral intent to adopt [22]. This is because they used the app, as we requested, to perform various tasks as part of the study. Therefore, in future studies, we would want to evaluate its usability to address users' experience issues and measure technology acceptance [22] to identify how to design for widescale adoption of an app to help people collaborate with their loved ones to manage mobile privacy and security. Lastly, the study design did not include a control condition, which means that any effects from the community oversight mechanism cannot be differentiated from changes that may have occurred through using the app, such as increased attention on app permissions and privacy and security. Therefore, the results cannot conclusively demonstrate a causal relationship between the usage of CO-oPS with communities and the dependent variables we analyzed. However, our qualitative insights provide evidence that some of the positive effects could be attributed to using the CO-oPS app. Moreover, there might be a survivorship bias effect in our results, as those who dropped out did not perceive any benefits to the app. Future research should investigate whether the same findings would hold for control groups and prevent potential survivorship bias.

8 Conclusion

Managing mobile privacy and security as an individual is hard. We believe community oversight is one potential social mechanism that can allow community members to exchange help regarding their mobile privacy and security decisions. Our CO-oPS app was developed to evaluate this idea of community oversight in building community collective efficacy for groups managing their mobile privacy and security together. Our results provide empirical evidence that community oversight can potentially have an impact on individuals and communities alike. Given the continued proliferation and adoption of smartphones and mobile apps, we believe apps that facilitate community oversight are an essential tool for communities to help one another keep their personal information safe and secure. We will continue to build upon this work to examine how we can help people successfully co-manage mobile privacy and security within their communities.

Acknowledgments

We acknowledge the contributions of Nikko Osaka, Anoosh Hari, and Ricardo Mangandi, in the CO-oPS app development. We would also like to thank the individuals who participated in our study. This research was supported by the U.S. National Science Foundation under grants CNS-1814068, CNS-1814110, and CNS-2326901. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the U.S. National Science Foundation.

References

[1] Zainab Agha, Karla Badillo-Urquiola, and Pamela J. Wisniewski. "strike at the root": Co-designing real-time social media interventions for adolescent online risk prevention. *Proc. ACM Hum.-Comput. Interact.*, 7(CSCW1), apr 2023.

[2] Mamtaj Akter, Leena Alghamdi, Dylan Gillespie, Nazmus Sakib Miazi, Jess Kropczynski, Heather Lipford, and Pamela J. Wisniewski. CO-OPS: A mobile app for community oversight of privacy and security. In *Companion Publication of the 2022 Conference on Computer Supported Cooperative Work and Social Computing*, CSCW'22 Companion, page 179–183, New York, NY, USA, 2022. Association for Computing Machinery.

[3] Mamtaj Akter, Leena Alghamdi, Jess Kropczynski, Heather Richter Lipford, and Pamela J. Wisniewski. It takes a village: A case for including extended family members in the joint oversight of family-based privacy and security for mobile smartphones. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI EA '23, New York, NY, USA, 2023. Association for Computing Machinery.

[4] Mamtaj Akter, Amy J. Godfrey, Jess Kropczynski, Heather R. Lipford, and Pamela J. Wisniewski. From parental control to joint family oversight: Can parents and teens manage mobile online safety and privacy as equals? *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW1), apr 2022.

[5] Zaina Aljallad, Wentao Guo, Chhaya Chouhan, Christy LaPerriere, Jess Kropczynski, Pamela Wisnewski, and Heather Lipford. Designing a Mobile Application to Support Social Processes for Privacy. In *Workshop on Usable Security, Internet Society*, 2019.

[6] Hazim Almuhimedi, Florian Schaub, Norman Sadeh, Idris Adjerid, Alessandro Acquisti, Joshua Gluck, Lorrie Faith Cranor, and Yuvraj Agarwal. Your Location has been Shared 5,398 Times! A Field Study on Mobile App Privacy Nudging. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 787–796, New York, NY, USA, April 2015. Association for Computing Machinery.

[7] Ashwaq Alsoubai, Reza Ghaiumy Anaraky, Yao Li, Xinru Page, Bart Knijnenburg, and Pamela J. Wisniewski. Permission vs. app limiters: Profiling smartphone users to understand differing strategies for mobile privacy management. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery.

[8] Monica Anderson. Mobile apps, privacy and permissions: 5 key takeaways, 2015.

[9] Mehrdad Bahrini, Nina Wenig, Marcel Meissner, Karsten Sohr, and Rainer Malaka. Happypermi: Presenting critical data flows in mobile application to raise user security awareness. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI EA '19, page 1–6, New York, NY, USA, 2019. Association for Computing Machinery.

[10] Albert Bandura. Self-efficacy mechanism in human agency. *American psychologist*, 37(2):122, 1982.

[11] Joseph Bonneau. Alice and Bob's life stories: Cryptographic communication using shared experiences. In *17th International Workshop on Security Protocols*, 2009.

[12] Joseph Bonneau, Jonathan Anderson, and Luke Church. Privacy suites: Shared privacy for social networks. In *Proceedings of the 5th Symposium on Usable Privacy and Security*, SOUPS '09, New York, NY, USA, 2009. Association for Computing Machinery.

[13] Paolo Calciati, Konstantin Kuznetsov, Alessandra Gorla, and Andreas Zeller. Automatically granted permissions in android apps: An empirical study on their prevalence and on the potential threats for privacy. In *Proceedings of the 17th International Conference on Mining Software Repositories*, MSR '20, page 114–124, New York, NY, USA, 2020. Association for Computing Machinery.

[14] J.M. Carroll and D.D. Reese. Community collective efficacy: structure and consequences of perceived capacities in the Blacksburg Electronic Village. In *36th Annual Hawaii International Conference on System Sciences, 2003. Proceedings of the*, pages 10 pp.–, Big Islane, HI, USA, January 2003. Institute of Electrical and Electronics Engineers.

[15] John M. Carroll, Mary Beth Rosson, and Jingying Zhou. Collective efficacy as a measure of community. In *Proceedings of the SIGCHI Conference on Human Factors*

in Computing Systems, CHI '05, page 1–10, New York, NY, USA, 2005. Association for Computing Machinery.

[16] Erika Chin, Adrienne Porter Felt, Vyas Sekar, and David Wagner. Measuring user confidence in smartphone security and privacy. In *Proceedings of the Eighth Symposium on Usable Privacy and Security*, SOUPS '12, New York, NY, USA, 2012. Association for Computing Machinery.

[17] Chhaya Chouhan, Christy M. LaPerriere, Zaina Aljallad, Jess Kropczynski, Heather Lipford, and Pamela J. Wisniewski. Co-designing for community oversight: Helping people make privacy and security decisions together. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), nov 2019.

[18] Lorrie Faith Cranor, Adam L. Durity, Abigail Marsh, and Blase Ur. Parents' and teens' perspectives on privacy in a technology-filled world. In *Proceedings of the Tenth USENIX Conference on Usable Privacy and Security*, SOUPS '14, page 19–35, USA, 2014. USENIX Association.

[19] Lee J. Cronbach. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3):297–334, September 1951.

[20] Sauvik Das, Tiffany Hyun-Jin Kim, Laura A. Dabbish, and Jason I. Hong. The effect of social influence on security sensitivity. In *Proceedings of the Tenth USENIX Conference on Usable Privacy and Security*, SOUPS '14, page 143–157, USA, 2014. USENIX Association.

[21] Sauvik Das, Adam D.I. Kramer, Laura A. Dabbish, and Jason I. Hong. The role of social influence in security feature adoption. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '15, page 1416–1426, New York, NY, USA, 2015. Association for Computing Machinery.

[22] Fred Davis. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13:319, 1989.

[23] Paul DiGioia and Paul Dourish. Social navigation as a model for usable security. In *Proceedings of the 2005 Symposium on Usable Privacy and Security*, SOUPS '05, page 101–108, New York, NY, USA, 2005. Association for Computing Machinery.

[24] Paul Dourish, Rebecca E. Grinter, Jessica Delgado de la Flor, and Melissa Joseph. Security in the wild: user strategies for managing security as an everyday, practical problem. *Personal and Ubiquitous Computing*, 8(6):391–401, November 2004.

[25] Mica R. Endsley. Toward a Theory of Situation Awareness in Dynamic Systems. *Human Factors*, 37(1):32–64, March 1995. Publisher: SAGE Publications Inc.

[26] Johannes Feichtner and Stefan Gruber. Understanding privacy awareness in android app descriptions using deep learning. In *Proceedings of the Tenth ACM Conference on Data and Application Security and Privacy*, pages 203–214, 2020.

[27] Adrienne Porter Felt, Elizabeth Ha, Serge Egelman, Ariel Haney, Erika Chin, and David Wagner. Android permissions: User attention, comprehension, and behavior. In *Proceedings of the Eighth Symposium on Usable Privacy and Security*, SOUPS '12, New York, NY, USA, 2012. Association for Computing Machinery.

[28] Jennifer Fereday and Eimear Muir-Cochrane. Demonstrating rigor using thematic analysis: A hybrid approach of inductive and deductive coding and theme development. *International Journal of Qualitative Methods*, 5(1):80–92, 2006.

[29] Denzil Ferreira, Vassilis Kostakos, Alastair R. Beresford, Janne Lindqvist, and Anind K. Dey. Securacy: An empirical investigation of android applications' network usage, privacy and security. In *Proceedings of the 8th ACM Conference on Security & Privacy in Wireless and Mobile Networks*, WiSec '15, New York, NY, USA, 2015. Association for Computing Machinery.

[30] Arup Kumar Ghosh, Karla Badillo-Urquiola, Shion Guha, Joseph J. LaViola Jr, and Pamela J. Wisniewski. Safety vs. surveillance: What children have to say about mobile apps for parental control. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–14, New York, NY, USA, 2018. Association for Computing Machinery.

[31] Jeremy Goecks and Elizabeth Mynatt. Supporting privacy management via community experience and expertise. *Communities and Technologies 2005*, 01 2005.

[32] Majid Hatamian. "hard to understand, easy to ignore:" an automated approach to predict mobile app permission requests: Student research abstract. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, SAC '19, page 1979–1982, New York, NY, USA, 2019. Association for Computing Machinery.

[33] Darren Hayes, Francesco Cappa, and Nhien An Le-Khac. An effective approach to mobile device management: Security and privacy issues associated with mobile applications. *Digital Business*, 1(1):100001, 2020.

[34] Qatrunnada Ismail, Tousif Ahmed, Kelly Caine, Apu Kapadia, and Michael K Reiter. To permit or not to

permit, that is the usability question: Crowdsourcing mobile apps' privacy permission settings. *Proc. Priv. Enhancing Technol.*, 2017(4):119–137, 2017.

[35] Patrick Gage Kelley, Sunny Consolvo, Lorrie Faith Cranor, Jaeyeon Jung, Norman Sadeh, and David Wetherall. A Conundrum of Permissions: Installing Applications on an Android Smartphone. In Jim Blyth, Sven Dietrich, and L. Jean Camp, editors, *Financial Cryptography and Data Security*, Lecture Notes in Computer Science, pages 68–79, Berlin, Heidelberg, 2012. Springer.

[36] Jess Kropczynski, Zaina Aljallad, Nathan Jeffrey Elrod, Heather Lipford, and Pamela J. Wisniewski. Towards building community collective efficacy for managing digital privacy and security within older adult communities. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW3), jan 2021.

[37] Jess Kropczynski, Reza Ghaiumy Anaraky, Mamtaj Akter, Amy J. Godfrey, Heather Lipford, and Pamela J. Wisniewski. Examining collaborative support for privacy and security in the broader context of tech caregiving. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW2), oct 2021.

[38] Jialiu Lin, Shahriyar Amini, Jason I. Hong, Norman Sadeh, Janne Lindqvist, and Joy Zhang. Expectation and purpose: Understanding users' mental models of mobile app privacy through crowdsourcing. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, UbiComp '12, page 501–510, New York, NY, USA, 2012. Association for Computing Machinery.

[39] Heather Richter Lipford and Mary Ellen Zurko. Someone to watch over me. In *Proceedings of the 2012 New Security Paradigms Workshop*, NSPW '12, page 67–76, New York, NY, USA, 2012. Association for Computing Machinery.

[40] Bin Liu, Mads Schaarup Andersen, Florian Schaub, Hazim Almuhimedi, Shikun (Aerin) Zhang, Norman Sadeh, Yuvraj Agarwal, and Alessandro Acquisti. Follow My Recommendations: A Personalized Privacy Assistant for Mobile App Permissions. pages 27–41, 2016.

[41] Rui Liu, Junbin Liang, Jiannong Cao, Kehuan Zhang, Wenyu Gao, Lei Yang, and Ruiyun Yu. Understanding mobile users' privacy expectations: A recommendation-based method through crowdsourcing. *IEEE Transactions on Services Computing*, 12(2):304–318, 2019.

[42] Haoran Lu, Luyi Xing, Yue Xiao, Yifan Zhang, Xiaojing Liao, XiaoFeng Wang, and Xueqiang Wang. Demystifying resource management risks in emerging mobile app-in-app ecosystems. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, CCS '20, pages 569–585. Association for Computing Machinery, 2020.

[43] Michael Lutaaya. Rethinking app permissions on ios. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI EA '18, page 1–6, New York, NY, USA, 2018. Association for Computing Machinery.

[44] Mary Madden. Privacy, Security, and Digital Inequality, September 2017. Publisher: Data & Society Research Institute.

[45] Tamir Mendel and Eran Toch. Susceptibility to Social Influence of Privacy Behaviors | Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, 2017.

[46] Tamir Mendel and Eran Toch. Social support for mobile security: Comparing close connections and community volunteers in a field experiment. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery.

[47] Eyitemi Moju-Igbene, Hanan Abdi, Alan Lu, and Sauvik Das. "how do you not lose friends?": Synthesizing a design space of social controls for securing shared digital resources via participatory design jams. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 881–898, Boston, MA, August 2022. USENIX Association.

[48] Savanthi Murthy, Karthik S. Bhat, Sauvik Das, and Neha Kumar. Individually vulnerable, collectively safe: The security and privacy practices of households with older adults. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1), apr 2021.

[49] Norbert Nthala and Ivan Flechais. Informal support networks: An investigation into home data security practices. In *Proceedings of the Fourteenth USENIX Conference on Usable Privacy and Security*, SOUPS '18, page 63–82, USA, 2018. USENIX Association.

[50] 1615 L. St NW, Suite 800 Wash., and DC 20036USA202-419-4300 | Main202-857-8562 | Fax202-419-4372 | Media Inquiries. Mobile Fact Sheet.

[51] Jinkyung Park, Eiman Ahmed, Hafiz Asif, Jaideep Vaidya, and Vivek Singh. Privacy Attitudes and COVID Symptom Tracking Apps: Understanding Active Boundary Management by Users. In Malte Smits, editor, *Information for a Better World: Shaping the Global Future*, pages 332–346, Cham, 2022. Springer International Publishing.

[52] Sai Teja Peddinti, Igor Bilogrevic, Nina Taft, Martin Pelikan, Úlfar Erlingsson, Pauline Anthonysamy, and Giles Hogben. Reducing permission requests in mobile apps. In *Proceedings of the Internet Measurement Conference*, IMC '19, page 259–266, New York, NY, USA, 2019. Association for Computing Machinery.

[53] Erika Shehan Poole, Marshini Chetty, Tom Morgan, Rebecca E. Grinter, and W. Keith Edwards. Computer help at home: methods and motivations for informal technical support. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pages 739–748, New York, NY, USA, April 2009. Association for Computing Machinery.

[54] Emilee Rader and Rick Wash. Identifying patterns in informal sources of security information. *Journal of Cybersecurity*, 1(1):121–144, September 2015. Publisher: Oxford Academic.

[55] Bahman Rashidi, Carol Fung, Anh Nguyen, Tam Vu, and Elisa Bertino. Android user privacy preserving through crowdsourcing. *IEEE Transactions on Information Forensics and Security*, 13(3):773–787, 2018.

[56] Joel Reardon, Álvaro Feal, Primal Wijesekera, Amit Elazari Bar On, Narseo Vallina-Rodriguez, and Serge Egelman. 50 ways to leak your data: An exploration of apps' circumvention of the android permissions system. In *WINTER 2019, VOL. 44, NO. 4*, pages 603–620, Boston, MA, United States, 2019. USENIX.

[57] Elissa M Redmiles, Amelia R Malone, and Michelle L Mazurek. I think they're trying to tell me something: Advice sources and selection for digital security. In *2016 IEEE Symposium on Security and Privacy (SP)*, pages 272–288. IEEE, 2016.

[58] Alireza Sadeghi, Reyhaneh Jabbarvand, Negar Ghorbani, Hamid Bagheri, and Sam Malek. A temporal permission analysis and enforcement framework for android. In *Proceedings of the 40th International Conference on Software Engineering*, ICSE '18, page 846–857, New York, NY, USA, 2018. Association for Computing Machinery.

[59] Seymour B Sarason. *The psychological sense of community: Prospects for a community psychology.* Jossey-Bass, 1974.

[60] Stuart Schechter and Joseph Bonneau. Learning assigned secrets for unlocking mobile devices. In *Eleventh Symposium On Usable Privacy and Security (SOUPS 2015)*, pages 277–295, Ottawa, July 2015. USENIX Association.

[61] Anna C Squicciarini, Heng Xu, and Xiaolong Zhang. Cope: Enabling collaborative privacy management in online social networks. *Journal of the American Society for Information Science and Technology*, 62(3):521–534, 2011.

[62] Sarina Till and Melissa Densmore. A characterization of digital native approaches to mobile privacy and security. In *Proceedings of the South African Institute of Computer Scientists and Information Technologists 2019*, SAICSIT '19, New York, NY, USA, 2019. Association for Computing Machinery.

[63] Emily A. Vogels and Monica Anderson. Americans and Digital Knowledge. *Pew Research*, October 2019.

[64] Hue Watson, Eyitemi Moju-Igbene, Akanksha Kumari, and Sauvik Das. "we hold each other accountable": Unpacking how social groups approach cybersecurity and privacy together. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–12, New York, NY, USA, 2020. Association for Computing Machinery.

[65] Pamela Wisniewski, Arup Kumar Ghosh, Heng Xu, Mary Beth Rosson, and John M. Carroll. Parental control vs. teen self-regulation: Is there a middle ground for mobile online safety? In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, page 51–69, New York, NY, USA, 2017. Association for Computing Machinery.

[66] Bo Zhang and Heng Xu. Privacy nudges for mobile applications: Effects on the creepiness emotion and privacy attitudes. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work; Social Computing*, CSCW '16, page 1676–1690, New York, NY, USA, 2016. Association for Computing Machinery.

[67] Hengshu Zhu, Hui Xiong, Yong Ge, and Enhong Chen. Mobile app recommendations with security and privacy awareness. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, page 951–960, New York, NY, USA, 2014. Association for Computing Machinery.

Appendix A Survey Scales

Community Oversight Model Constructs: (Derived from Chouhan et al.'s conceptual model of Community Oversight [17])
Transparency
1. The app gave me a transparent view into the apps installed and permissions granted on my own mobile device.
2. The app gave me a transparent view of the apps installed and permissions granted on the mobile devices of others.
3. The app gave us all a transparent view of the apps installed and permissions granted on the mobile devices of our community.
Awareness
1. The app made me more aware of my own mobile privacy and security decisions.
2. The app made me more aware of the mobile privacy and security decisions of others.
3. The app increased overall awareness of the mobile privacy and security decisions of our community as a whole.
Trust
1. The app helped me foster trust in the mobile privacy and security decisions of others in my community.
2. The app helped others in my community foster trust in my mobile privacy and security decisions.
3. The app helped foster trust in the mobile privacy and security decisions of our community as a whole.
Individual Participation
1. The app helped me make privacy and security decisions for myself.
2. The app helped me be involved in others' privacy and security decisions.
3. The app helped individuals in the community participate in privacy and security decisions of our community.
Community Participation
1. The app enabled me to participate in a community that helps one another regarding our mobile privacy and security decisions.
2. The app enabled others to participate in a community that helps one another regarding our mobile privacy and security decisions.
3. The app enabled the community to help one another regarding our mobile privacy and security decisions.

Community Trust (Derived from Chouhan et al.'s conceptual model of Community Oversight [17])
1. I trust others in my community to protect my private information.
2. I trust others in my community to give me advice about mobile privacy and security.
3. Others in my community trust me to protect their private information.
4. Others in my community trust me to give them advice about mobile privacy and security.

Community Belonging (Pre-validated by Carroll et al. [14] and Sarason et al. [59])
1. I can get what I need in this community.
2. This community helps me fulfill my needs.
3. I feel like a member of this community.
4. I belong in this community.
5. I have a say about what goes on in this community.
6. People in this community are good at influencing each another.
7. I feel connected to this community.
8. I have a good bond with others in this community.

Self-Efficacy (Pre-validated by Kropzynski et al. [36] based on a modified version from Bandura [10])
1. I know that if I worked hard to learn about mobile privacy and security, I could make good decisions.
2. Mobile privacy and security decision-making is not too complicated for me to understand.
3. I think I am the kind of person who would learn to use best practices for good mobile privacy and security decision-making.
4. I think I am capable of learning to help others make good mobile privacy and security decisions.
5. Given a little time and training, I know I could learn about best practices for good mobile privacy and security decision-making for myself and my community.

Community-Collective Efficacy (Pre-validated by Kropzynski et al. [36] based on a modified version from Carroll et al. [15])
1. Our community can cooperate to improve the quality of our decisions about mobile privacy and security.
2. Despite other obligations, we can find time to discuss our decisions about mobile privacy and security.
3. As a community, we can handle the mistakes and setbacks resulting from our decisions about mobile privacy and security without getting discouraged.

4. I am confident that we can be united in the decisions we make about mobile privacy and security that we present to outsiders.
5. As a community, we provide care and help for one another regarding our mobile privacy and security decisions.
6. Our community can leverage outside resources and services for our members to ensure the quality of mobile privacy and security decisions.
7. Our community can provide information for people with different interests and needs when it comes to mobile privacy and security decision-making.

Appendix B Sample Questions of Followup Interview

- *Prior to participating in this study, how did you decide which apps are safe or unsafe to install on your mobile devices?*
- *How did you decide whether to accept or deny a permission request for an app?*
- *Did you ever review the permission lists of the apps installed on your phone? Why or why not? How?*
- *How frequently did people in your community discuss mobile privacy and security issues with one another?*
- *During the study, how frequently did your community members discuss mobile privacy and security decisions with one another?*
- *During the study, how did you communicate with others who were part of your community?*
- *During the study, how did you manage your mobile privacy and security decisions? Did you see any changes compared to prior to the study? Why or why not?*
- *Can you explain how and why the app did or did not help provide transparency into the mobile privacy and security decisions of other people in your community?*
- *How and why did the app or did not help raise awareness in your community about mobile privacy and security?*
- *How and why did the app or did not enable you and individuals in your community to provide feedback and guidance about others' mobile privacy and security?*
- *How and why did the app or did not help you work together as a community about mobile privacy and security?*
- *Were there any problems or concerns you or others in your community encountered when using the app?*
- *If given the option, would you want to continue using the CO-oPS app after this study? Why or why not?*
- *Who do you think would be benefited the least from using the app and why? Who would be most benefited and why?*
- *Is there anything else that you would have liked the app to do? Any changes you would have liked on how the app currently works?*

Appendix C Codebook

Table A.1: Codebook

Codes	Illustrative Quotations
Transparency	
Visibility to own S&P (security and privacy) (76%, N=39)	*"So actually using co-ops, like, for me, I got to see the list, like, what the apps, what the actual permission all the apps are using and like, what access they have. Like the list of all at the same place. For me, it was like, good to have this."* -C11P3
Visibility to others S&P (67%, N=34)	*"I think having this app actually made me more, see these things of others, because it made it easier now to check, not only yours, but also other people's security settings."* -C18P1
Violation of own privacy when other's view (38%, N=20)	*"As some of the community are not someone who not much close, I wasnt that much confident when it came to share my apps and show my things, you know."* -C08P2
Violation of own privacy when forget to hide (27%, N=14)	*"I didn't want to show a few apps, to my community members, but, as CO-oPS crashed the first time and I had to reinstall it. Then, I forgot to hide those apps. And so I think that is a privacy issue, which most people won't like it."* -C11P1
Violation of others' privacy (25%, N=13)	*"While using the app, my friends and I discussed privacy more than security because we can see the apps on our friend phones and I think that's very not a good thing. I did not feel good."*-C11P2
Defeats the purpose (25%, N=13)	*"But sometimes, so while people are using some apps and keeping it private to them, they dont share with anyone but yeah, then I think this app wont help much for anyone"* -C06P2
Awareness	
Overall S&P awareness (94%, N=48)	*"Sometimes we allow some permissions without understanding what's been packed. So after exploring that CO-oPS, I usually get to think twice about my apps, which really cool, I am more concerned about whether to allow or not allow any permissions to secure your phones. I would say it's very helpful to change my mind. And it helped me to be more careful about my mobile security."* -C15P1

Continued on next page

Table A.1 – continued from previous page

Codes	Illustrative Quotations
Compare own S&P with others (86%, N=44)	*"I think this is a great feature. Because with this, you are able to see and compare like, if what you are using and what others are using, it is like comparable Or you can just know what you are doing others are not. I guess you can help yourself."* -C17P1
Keep track of own S&P (82%, N=42)	*"Earlier I couldn't know about what is there and what is not because I thought I had few apps the apps I did install. Then here [on CO-oPS app] I see I have more apps that I did not see it before... I think it helps, it feels like gives you to see what do you have on the phone, and the stuff that are accessed by the apps."* -C08P1
Aware of others' S&P (57%, N=29)	*"So like seeing the option of like, every single app, and then seeing like what's granted and denied, that definitely helped a lot to see what each member, what apps did they have, and also what like permissions they grant. So it helps me realize what they're granting or not granting, so that I need to I help them or not."* -C02P4
Aware of community's S&P Changes (39%, N=20)	*"One of the benefits of it is, on the community section, I can go through my friend's app changes, which permissions of which apps you changed. And I can go ahead and do that and change it and have fun. Okay."* -C11P1
Increased awareness from pro tips (39%, N=20)	*"So on this pro tip section in where you can know the basic information, like basic knowledge that you can just learn from and become careful about the app settings... I think this section talked some senses in us."* -C06P2
Doesnt inform community about Apps Changes (18%, N=9)	*"So, you see in the community, we get to know about the changes for the permissions, but we do not get any community posts for the app installing or installing. I think this is also important. When someone gets rid of an app, everuyone should know, right."* -C15P1
Trust	
Trust others' advice (51%, N=26)	*"[Name] let me reconsider what I am doing, because when he tells me warns me, you are more likely to take it seriously. It'll come to light in your mind for sure. Yeah, I did change some of the things, yeah I think he was right. I see the stuff he warns me about are all good."* -C11P2
Less aware community members (49%, N=25)	*"I dont think they were much of aware. They do not care of all this, you know, privacy and security stuff, so I am not sure they used it much."*-C11P3
Less tech-savvy community members (37%, N=19)	*"For example, my mom... whenever she goes to the Facebook or YouTube, she asks questions. So, she cant be able to understand these privacy and security, its just so beyond her capacity. So I doubt she would be someone to rely on."*-C16P1
Individual Participation	
Made own S&P changes (67%, N=34)	*"I got rid of some of my permissions. I haven't really thought of that before. Right now it has come to my knowledge that yes, it is a big problem and even scary. But I have that control, if you know what permissions are problematic, and what are necessary, you can always try clean up. Now cleaning up my phone has become a bit of a priority to me."* -C02P1
Provided feedback (41%, N=21)	*"I reviewed X's mobile privacy, I saw he was giving a permission, don't remember which one, then I told him that, allowing that permission is not good. And then I gave some good reasons why this is important to change this or not.* -C17P1
Less tech-savvy (41%, N=21)	*"I dont think anyone needed my advice. I know they are careful, much careful than I am because they all are very savvy."* -C18P2
Others are not close (39%, N=20)	*I dont think I did much. . . I would be interested to help someone when I care them, maybe my parents mostly."*-C14P2
Fewer app users (29%, N=15)	*"I did not use it much. I'm a very minimalist in my apps. So at this point, the apps that I have, I know what I have. My advice to others is use minimal apps and make your life easy.."*-C03P5
Community Participation	
Learned from community (78%, N=40)	*"We could review each other's permissions and we could Share, so we could be careful about our privacy and things. And having your community's apps and permission in CO-oPS, you can just learn by yourself like maybe you don't really have to grant this permission."* -C18P2
Increased discussion in community (71%, N=36)	*"We had frequent discussions when we had discussions about what kind of security and permissions we have or on each other's phone, or in general the security issues out there. And I think the other day when we met, we were giving away some information. I think we also mentioned some of our apps are taking unnecessary data. For those apps purpose, the permissions were not necessary. So we asked to turn it off. And I don't know if they did change that, but I did. But yeah, that kind of interaction truly happened among us. And we had we shared opinion and try to suggest each other that this is not right."* -C06P4
Received feedback from community (53%, N=27)	*"One of the action items we had a task like look through permissions and tell them like, hey, like, maybe you shouldn't do it. I think I received a message from X like, Hey, you have Bose like music app has access to your GPS location for some reason. Oh, wow. Which I did not notice it before. This was like, I really thanked him."* -C09P2
Less tech-savvy community (31%, N=16)	*"I think when your community is not tech savvy, they wont feel the importance of this security and privacy. I can see to be an effective community at least some people must be tech savvy so that they can educate everyone else."* -C16P1
Tech savvy community (27%, N=14)	*"We didnt find it useful, not really, because my community is from computer science background. I think we are already aware of these things. So, we dont need others advice."* -C11P5
Community Trust and Belonging	
Good relationship with community (86%, N=44)	*"We live in a same community, so we have a very good relation with the other people like X and all the other four members because we almost live in very close to and very similar minded community. So and I have personally good relationship with X that also drives me to participate in this research. So yeah, we try to go outing and explore things together."* -C15P1

Continued on next page

Table A.1 – continued from previous page

Codes	Illustrative Quotations
Trusted others to keep S&P Info Private (47%, N=24)	*"I guess, like the thought that they are my close circles. Like I know sharing my apps with them is safe."* -C15P1
Depend on the community for S&P (24%, N=12)	*"With X in my group at least I know that if he saw something he didn't thought wasn't proper, he will definitely let us know, let me and my husband know.. We have that kind of relation, he, Yeah, he would let us know and he would tell us this just to delete that, we have that kind of relationship so, we know we can trust him, We know that,"* -C02P1
Had security concern for sharing S&P (18%, N=9)	*"I have my Chase app, if someone on the family, like my sons, know I have this app and can somehow get my phone,... if the app is logged in already, they can just transfer the money immediately."*-C02P3
Self Efficacy	
Gained confidence in S&P (80%, N=41)	*"Okay, so, I will say that what is the purpose of this app? Like if it is like Facebook or WhatsApp, then it will use my contacts, my contact information can use or my photos they can use. But why they should go to my phone call manage permission or there will track my other applications permission. That doesn't make sense. So it conflicts with the purpose of this application. See, this is new. I can now differentiate whats necessary or what not."* -C10P1
Now know how to change permissions (57%, N=29)	*"So I actually now can use the settings to go directly change the permissions. Its much easier now. It has become like I randomly go check some apps and do changes instantly if I feel like."* -C12P2
Already confident in S&P (31%, N=16)	*"I would say that'd be me. I'm pretty knowledgeable regarding, you know, the whole privacy and phones, I try to be secure about my own apps. Yeah, I think I am very careful with permissions and such. I know how to change things."* -C22P1
Community Collective Efficacy	
Felt teamwork for S&P (88%, N=45)	*"I mean, offline, or virtually, we kind of worked together, we talked, we get each other's knowledge. We could easily just start a discussion about any apps and permissions stuff... I will say it kind of, we work together in this."* -C17P1
Reached out to community (67%, N=34)	*" think one thing is that I'm a little more confident of it now. So, when I'm giving permissions, I now can tell that could be the things that needed for a discussion. I do go to [Name] to ask what he thinks would do. what he thinks if the permission is needed or not needed for the app. I do my permissions like this now."* -C03P5

Appendix D Descriptive Statistics of Community Oversight Construct Items

Table A.2: Wilcoxon Rank-Sum Tests of Pre and Post-Study Responses to the Community Oversight Construct Items

Constructs	Pre-Study				Post-Study					
	M1	*SD1*	*S1*	*K1*	*M2*	*SD2*	*S2*	*K2*	*V-val*	*p-val*
Transparency:										
The app gave me transparency of my apps and permissions	4.18	0.86	-0.89	0.73	4.40	0.72	-0.79	-0.68	716	0.08
The app gave me transparency in others' apps and permissions	4.02	0.90	-0.85	0.97	4.32	0.77	-0.91	0.29	696*	**0.026**
The app gave us transparency in whole community's app	3.98	0.99	-0.83	0.30	4.30	0.76	-0.86	-0.27	654.5*	**0.012**
Awareness:										
The app made me aware of my S&P decisions	4.12	0.91	-1.31	2.53	4.49	0.67	-1.15	1.10	462.5**	**0.001**
The app made me aware of others' S&P decisions	3.81	0.99	-0.74	0.42	4.28	0.85	-1.01	0.31	804**	**0.001**
The app made me aware of whole community's S&P decisions	3.91	0.94	-0.69	0.09	4.23	0.84	-1.01	0.52	880.5***	**<0.001**
Trust:										
The app helped me foster trust in others' S&P decisions	3.65	0.98	-0.37	-0.21	4.19	0.81	-0.74	-0.00	405***	**<0.001**
The app helped others foster trust in my S&P decisions	3.57	0.91	-0.16	-0.32	4.02	1.02	-0.96	0.39	546.5***	**<0.001**
The app helped foster trust in whole community's S&P decisions	3.60	0.84	-0.43	0.17	4.12	0.94	-1.09	1.15	467***	**<0.001**
Individual Participation:										
The app helped me make S&P decisions for myself	3.96	0.98	-0.82	0.39	4.37	0.79	-1.43	2.84	727.5**	**0.002**
The app helped me involve in others' S&P decisions	3.62	1.02	-0.58	-0.13	4.17	0.92	-1.10	0.94	598***	**<0.001**
The app helped us involve in whole community's S&P decisions	3.75	1.00	-0.58	-0.13	4.16	0.82	-0.66	-0.23	522***	**<0.001**
Community Participation:										
The app enabled me to participate in community's S&P	3.85	0.90	-0.60	0.18	4.12	0.87	-0.91	0.82	760*	**0.021**
The app enabled others to participate in community's S&P	3.87	0.84	-0.52	0.42	4.17	0.91	-0.87	-0.10	771.5*	**0.011**
The app enabled whole community to participate in everyone's S&P	3.85	0.90	-0.69	0.32	4.25	0.89	-0.99	0.12	701**	**0.003**
Community Trust:										
I trust others to protect my information	3.99	0.28	-1.18	0.87	4.15	0.23	-1.21	1.62	1012.5	0.166
I trust others to give me advice	4.13	0.24	-1.10	1.10	4.23	0.20	-1.23	2.44	967.5	0.356
Others trust me to protect their information	4.04	0.24	-0.81	0.10	4.28	0.19	-0.49	-1.02	651*	**0.016**
Others trust me to give them advice	3.96	0.25	-0.72	0.11	4.18	0.22	-1.17	2.04	856*	**0.023**

*p<.05; **p<.01; ***p <.001

Data Privacy and Pluralistic Ignorance

Emilee Rader
Michigan State University
emilee@msu.edu

Abstract

This paper presents the results of an online survey experiment with 746 participants that investigated whether social norms influence people's choices about using technologies that can infer information they might not want to disclose. The results show both correlational and causal evidence that empirical expectations (beliefs about what others do) and normative expectations (beliefs about what others believe) influence choices to use mobile devices in ways that generate data that could be used to make sensitive inferences. However, participants also reported concern about data privacy, and lower behavioral intentions for vignettes involving more invasive inferences. Pluralistic ignorance is a phenomenon where individuals behave in ways they privately disagree with, because they see others around them behaving the same way and assume this is evidence most people approve of the behavior. These results are consistent with the existence of pluralistic ignorance related to data privacy, and suggest that interventions focused on transparency about data practices are not enough to encourage people to make different privacy choices.

1 Introduction

Every time someone decides to use their mobile device to set an alarm, send a text message, look up directions, or any other action, they are making a choice that has privacy implications, even if they are not explicitly aware of it. Using many technologies for normal, everyday purposes generates data that supports making inferences about a user's body, activities and personal characteristics that are difficult to anticipate and can be surprising, unsettling or harmful when used for unexpected purposes [10, 19, 21].

Many conceptualizations of privacy treat it as an individual right, which means that individuals are responsible for controlling their own information according to their concerns and preferences [24]. With respect to data privacy, this perspective is codified in the logic of "privacy self-management", or notice and choice [32], where privacy is a transaction between the technology user and the platform, system, or organization that is on the receiving end of the data. This individualistic framework limits the potential avenues for influencing people's data privacy decisions to individual-level approaches: increasing knowledge about the implications of making different choices, providing more fine-grained controls and widgets for expressing privacy preferences, and disclosing information about the collecting party's data practices.

But, privacy is inherently social [23]. It is well established that social norms play an important role in interpersonal disclosure decisions [17]. For example, in interpersonal privacy, disclosure rules and boundaries are often norm-based. People learn about appropriate and inappropriate disclosure behavior from others in their family or organizations they belong to, and subsequently form beliefs about what private information looks like and how it should be managed [18].

There is also evidence that behaviors that promote data privacy are subject to social judgments and influence. For example, previous research has found that using encryption to protect one's communications is perceived to be a behavior that makes one seem paranoid [8, 33], a stigmatized delusional state characterized by extreme suspicion [22]. And, Solove wrote that the "nothing to hide" argument is an extremely common response to finding out about unwanted data collection (e.g., "I'm not doing anything wrong, so I have nothing to hide"). This argument implies an assumption that only bad actors have reasons to want to protect their data privacy [31], so wanting privacy means one must be a bad actor. These examples illustrate that wanting data privacy is something that could cause a person to be judged negatively by others. However, people still say they value data privacy [20].

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.
USENIX Symposium on Usable Privacy and Security (SOUPS) 2023.
August 6–8, 2023, Anaheim, CA, USA

A norm-based phenomenon called pluralistic ignorance occurs when people engage in a behavior they privately do not believe in or approve of, but they do it anyway because they believe that everyone else approves of it [14]. With respect to data privacy, this could look like privately being concerned about protecting data about oneself, but choosing to use technologies that can generate invasive inferences due to social expectations. If data privacy decisions are subject to this type of social influence, it could mean that interventions that are intended to help people manage their data privacy but are based on individualistic assumptions would fail for reasons that would be hard to identify at the individual level.

This paper presents an experiment investigating whether social norms that conflict with personal privacy preferences influence technology use decisions that have data privacy implications. The results show that social expectations do influence choices to use potentially invasive technologies, despite participants' private concern about data privacy. These results support an interpretation that pluralistic ignorance exists related to data privacy, and suggest that awareness interventions intended to change people's behavior by increasing their knowledge about threats to privacy may be ineffective. This paper contributes novel results to the research literature about social influences on data privacy by showing through a controlled experiment that people may use technologies they feel privacy concern about because of their beliefs about others' approval or disapproval of those technologies.

2 Related Work

2.1 Social Influences on Privacy Choices

Recent research has focused on the idea that information about the behavior and choices of others may be helpful for people faced with making privacy and security decisions. For example, through a participatory design study Chouhan et al. [5] evaluated a mechanism to help people seek security and privacy advice from their community, and found that participants felt like the necessary expertise to help them make decisions did not exist in their circle of close family and friends. Nissen et al. [16] explored participants' reactions to the idea of delegating consent decisions to third parties, and found that trust in the expertise of the third party was an important factor in whether they would delegate or not. Naeini et al. [15] investigated the influence of information about others' privacy choices on participants' choices, and found that information from friends and experts had different effects—the most influential social cues occurred in scenarios where friends denied data collection, and experts allowed it. And, Krsek et al. [13] found that being shown suggestions for security and privacy settings from unknown experts and members of the public influenced participants to self-report that they would choose settings resembling what had been suggested to them. These papers share a common focus on providing social input to specific security and privacy decisions, through providing information about the experiences and behavior of others.

In contrast, the focus of this paper is on the potential that social norms might implicitly influence participants' willingness to use technologies that collect data about them and are capable of making invasive inferences. This paper explores, in a broader sense, whether the influence of social norms may help explain why people continue to use technologies that are bad for privacy, even while they say privacy is important to them.

2.2 Theoretical Background

Research on norms in social psychology focuses on what are referred to as descriptive and injunctive norms. Descriptive norms are based on observing the behavior of others and using that as an example of what one should do [11]. Injunctive norms refer to behaviors that are either reinforced or discouraged through feedback from other people regarding their approval or disapproval of the behavior [6].

It can be difficult to tell whether a collective behavior—one that is observed among many members of a group or community—is caused by one's beliefs about the behavior or beliefs of others. For example, if everyone outside is using an umbrella it may simply be because it is raining and they don't want to get wet. But, there may be a social norm that umbrellas are more acceptable than raincoats in that situation, and it is impossible to know whether the choice to use an umbrella is a result of an individual's personal preference or due to their beliefs about what others would think about their choice of rain gear.

Bicchieri [1] argues that collective behaviors can be independent or interdependent. Independent but similar behaviors arise due to situational factors, whereas interdependent behaviors arise due to social influences. Those social influences can be empirical (e.g., using an umbrella and not a raincoat because that's what one sees others doing) or normative (e.g., using an umbrella because one believes people would think someone who uses a raincoat instead of an umbrella is weird).

A key concept in assigning causation for a collective practice is the type of beliefs that guide behavior. If a group of people behaves in the same way coincidentally, that behavior is not influenced by a social norm. Therefore, we can identify that a collective behavior results from a social norm and is not just coincidental if individuals engage in the behavior because they believe it is commonly done by others, or if they do it because they believe that others approve of the behavior. Bicchieri [1] describes these two types of social influence this way:

- *Empirical expectations* are beliefs about how most others will behave in similar situations, and depend on observing others' behavior (similar to descriptive norms)

- *Normative expectations* are beliefs about what others approve/disapprove of in similar situations (similar injunctive norms)

Pluralistic ignorance is a situation in which people's beliefs about what others approve/disapprove of are incorrect. It occurs where there is a common behavior that people engage in because they see others doing it and believe that this is evidence that they all approve of it. But privately, in fact, most people dislike or disagree with the behavior. Pluralistic ignorance has been implicated in social phenomena as varied as the bystander effect [26], campus alcohol abuse [7], and climate change inaction [12]. Pluralistic ignorance is a visibility problem, where the information people have access to about others' behaviors leads to incorrect assumptions about their beliefs [28]. In the context of climate change, this would look like seeing most people around you driving gasoline engine pickup trucks and assuming that you're the only one who cares about greenhouse gas emissions [12].

A characteristic of pluralistic ignorance is that people believe that they know others' private opinions, but are actually incorrect about what those opinions are [7]. This is recognizable in the discourse about data privacy as the belief that nobody cares about privacy anymore, or that privacy is dead [25], when in fact people do care about privacy [30]. In the data privacy context, this could look like seeing others around you using always-on voice assistants and assuming they must not care about privacy, because if they did they wouldn't use them. It is important to discover whether people's data privacy decisions are affected by pluralistic ignorance, because this would help researchers and practitioners understand what kind of informational interventions would be likely to make a difference. For example, individualistic interventions focused on knowledge about data practices and privacy harms would be less effective in a situation where people's data privacy choices depend on normative assumptions about the behavior and beliefs of others.

2.3 Research Questions

For pluralistic ignorance to exist related to a behavior that an individual engages in, three things must be true. First, individuals have to perceive that the behavior is common among other people. In other words, there has to be an empirical expectation—a belief about what others do—that supports the behavior. Second, individuals have to believe that the behavior is something others approve of. There has to be a normative expectation—a belief about what others believe—that also supports the behavior. And third, individuals have to have a personal expectation—that is, their own, private belief—disliking the behavior, even if they do it (or are likely to do it) anyway themselves. Investigating these three conditions that amount to pluralistic ignorance is the purpose of this study, and each one has an associated research question.

First, the study investigates the relationship between participants' existing empirical and normative expectations and their use of their mobile device in a context that could produce unwanted inferences. The first research question asks whether a person's beliefs about what others do (empirical expectations) and beliefs about what others believe (normative expectations) influence an actual privacy-related behavior.

> RQ1: Is there a relationship between self-reported empirical and/or normative expectations and using a technology that has privacy implications?

Next, the study uses a vignette about a hypothetical mobile device user, similar to the participant, to investigate the influence of empirical and normative expectations that are experimentally manipulated via the vignette on compliance with the use of a mobile device that can make potentially invasive inferences. In other words, the experiment investigates whether norms have a causal influence on the likelihood of complying with the behavior described in the vignette. Using a vignette reduces the impact of social desirability bias, and makes it possible to ask about situations that may contradict the situation in the participant's real life.

> RQ2: Do empirical and normative expectations affect likelihood of compliance with a behavior that produces potentially invasive, unwanted inferences?

Because the vignette includes information about a possible inference that can result from using the mobile device as described, it acts as a kind of awareness intervention explicitly informing participants about this possibility. The third research question asks whether there is a relationship between this intervention and participants' behavioral intentions after learning about the inference.

> RQ3: Is there a relationship between the technology use context, including possible inferences, and behavioral intentions?

3 Method

A survey-based experiment was conducted with 746 participants, hosted on the Qualtrics platform. Data collection took place online during September 20-30, 2021. The Institutional Review Board which oversaw this experiment determined the research to be exempt.

3.1 Vignettes

The experiment involved presenting a very short text vignette (M = 62 words) to participants which described a hypothetical situation. The vignettes each began with the statement,

"Please imagine the following situation and answer the question that follows: Somebody like you lives in a very similar area of the country." The vignettes described the use of a mobile device that collects some type of data and makes inferences about the main character of the vignette, described as someone similar to the participant. The vignettes manipulated the context of the situation, information about the extent to which others use their mobile devices as described (the empirical expectation, a description of what others do), and whether others believe is is OK for people to use their mobile devices in that way (the normative expectation, a description of what others approve of). Details about characteristics of the main character of the vignette were deliberately left vague, to allow the participant to imagine someone similar to them in the ways that were most important or relevant to each individual participant.

The vignettes and experiment design were based on a study by Bicchieri et al. about normative influences on masking and social distancing during the early days of the COVID-19 pandemic [2]. They wrote that it can be difficult to measure the influence of norms on behavior via a survey asking participants about their own behavior, because self-report responses can be affected by social desirability bias. This means that participants' answers may reflect normative beliefs about how one should behave, rather than how the participant thinks they would behave in the situation. There is some evidence in prior work that privacy-preserving behaviors are labeled by others as paranoid or crazy [8, 20, 33], so social desirability bias could be a real problem in this research. By making the vignette about someone else, participants' responses are about others' behavior, not their own. Therefore, they may be willing to answer in a less biased way.

In addition, the vignettes are deliberately simple. The goal of this study was to investigate whether the behavior in the vignette, as imagined by the participant, is subject to social influence. For this study, it does not matter if each participant understands the technology or the vignette behavior slightly differently. The focus of the study is whether there is a social component (empirical or normative expectation) that influences expected compliance with the behavior.

3.2 Experiment Conditions

The experiment had three categorical independent variables: context (3 levels) x empirical expectations (2 levels) x normative expectations (2 levels). It used a full factorial design for a total of 12 between-subjects conditions. Participants were each assigned at random to one of the twelve conditions.

The context dimension refers to the description in the vignette of how the mobile device would be used, and inferences that would be possible due to this use. Three different contexts were used, because previous research has established that privacy is contextual, and privacy-related choices and behaviors depend on context [17]. The contexts in this experiment were based on scenarios from Rader [20], an interview study that investigated participants' reactions to hypothetical scenarios involving unexpected inferences made from sensor-based technologies. The contexts used in this experiment are as follows:

- The *alarm* context focused on using a mobile device as a wake-up alarm. The inference presented in the vignette was that the system could detect how often the user snoozes or sleeps through the alarm. This context was selected because it is a common use case for mobile devices, and an inference that participants in Rader's interview study [20] viewed as directly related to the purpose as a wake-up alarm. They also felt it could be seen as helpful information for changing one's sleep routine or habits.

- The *cookbook* context involved using one's mobile device as a digital cookbook. The inference presented in the vignette was that the system could analyze the foods the user likes to eat and determine how healthy the user is. This context was chosen because keeping recipes on a mobile device is something that is currently possible, but the inference about the user's health is not directly related to the purpose as a cookbook. Participants in Rader's study [20] appreciated the idea of being able to hands-free cooking or possible suggested ingredient substitutions or recipe recommendations to encourage healthier eating habits, but were concerned about unwanted inferences affecting their health insurance or otherwise indicating that they were being judged or evaluated as unhealthy because of the foods they eat.

- The *location* context involved allowing one's mobile device to collect location data that could be used to infer how often the user visits the restroom. This context was chosen because most mobile users allow location data to be collected by apps or their mobile operating system. However, the inference is not tied to a specific purpose for using the mobile device, and is something people would may be uncomfortable with because it violates a taboo about sharing information about one's bathroom behavior. In Rader's study, while some participants imagined that this inference could be useful for one's doctor if the goal was to collect data on a medical condition, nearly all participants had very strong, negative reactions to the idea of a mobile device making inferences about their bathroom habits.

The empirical expectation dimension refers to information in the vignette about how common it is that other residents in the hypothetical community use the mobile device for the purpose described in the vignette. The normative expectation dimension refers to information about how common it is that others approve of using the mobile device for that purpose. In

other words, the vignettes provided information about what other people do in the hypothetical situation (empirical expectation), and also about what other people believe should be done in that situation (normative expectation). Empirical and normative expectations each had two levels, "most" versus "few" other people. An example vignette is shown below, for the alarm (context) x high (empirical expectations) x high (normative expectations) condition. The text of the vignette closely follows the scenarios used in Rader's study [20]. See the replication materials, available online, for the full text of the 12 vignettes used in the experiment.

> Somebody like you lives in a very similar area of the country. **Most / Few** [empirical expectation][1] residents are using their mobile device as their wake-up alarm, which means it is possible for the system to detect how often they snooze or sleep through the alarm. **Most / Few** [normative expectation] residents also believe that it is OK for people to use their mobile device as their wake-up alarm.

3.3 Participants

Participants were recruited using the Qualtrics panel service. Eligible participants were mobile device users 18 years old or older who lived in the United States, and who had not had formal training or worked in a high-tech related field or discipline. The experiment used quotas for age (4.7% 18-20 years old, 41.3% 21-44 years old, 32.9% 45-64 years old, 21.1% 65+ years old) and gender (51% women) based on the 2019 U.S. Census Bureau Current Population Survey, Annual Social and Economic Supplement[2].

2539 participants started the survey by viewing the consent form. 194 declined consent, and 1523 were determined to be ineligible based on their answers to the screening questions. Eight additional participants were excluded when they did not agree to a quality commitment question. Finally, 64 responses were excluded before finishing the experiment where participants reported having "Good" or "Full" familiarity with a made-up word, and 4 more were excluded for answering all of the questions in less than 2 minutes. The final dataset for analysis includes 746 participants who completed the experiment. Participants ranged in age from 18 to 93 (M = 48, SD = 18). 50.7% were women, and 80% reported "White" as one of the ethnicity categories that described them. See the table in the Appendix for additional demographic details about the participants.

On average, it took participants 8 minutes to complete the experiment (SD = 6 min). They were allowed up to 24 hours to finish from the time they started reading the consent form. They received an incentive for completing the experiment in the form of gift cards or in-app credits equivalent to about $2 USD. This amount was determined by representatives of the Qualtrics panel service.

3.4 Procedure

Potential participants received a study invitation via an email message, and clicked on a link that directed them to the online survey. They first viewed the consent form, and after consenting were directed to a series of screening questions to determine their eligibility to participate. Eligible participants were assigned to one of 12 experiment conditions using a random number generator built in to the Qualtrics platform, and subsequently saw and answered questions about only one vignette.

The target size for each condition was 60 participants. The actual number of participants in each condition ranged from 54 to 72 (M = 62, SD = 5.8). The unequal n across conditions resulted from the method of random assignment, and a small number of participants that were excluded after assignment for data quality reasons (i.e., attention check, speeding through the survey). There was no correlation between the number of participants per condition and the number excluded in each condition. See Table 1 for how many participants were assigned to each condition, and the number per condition that were excluded after assignment.

Next, participants were asked a set of 7 questions that varied based on the experiment condition the participant was assigned to. These survey questions were designed based on the Bicchieri et al. study [2], as were the vignettes. Two questions first asked about participants' own past behavior and their beliefs about others' behavior, related to the context:

- *personal behavior*: "Do you [use your mobile device as your wake-up alarm | use an app on your mobile device as a digital cookbook | allow your mobile device to track your location]?" (Yes, No)

- *personal beliefs*: "Do you believe that it is OK for people to [use their mobile device as their wake-up alarm | use an app on their mobile device as a digital cookbook | allow their mobile device to track their location]?" (Yes, No)

An additional two questions asked about their perception of norms related to the context, in the form of empirical expectations about the behavior of others and the prevalence of believing it is OK to behave that way:

- *perceived empirical expectations*: "Please estimate the percentage of fellow residents in your area who [use their mobile device as their wake-up alarm | use an app on their mobile device as a digital cookbook | allow their mobile device to track their location]." (0-100 in increments of 10)

[1] The text in brackets was not presented to participants.

[2] See https://www2.census.gov/programs-surveys/demo/tables/age-and-sex/2019/age-sex-composition/2019gender_table1.xlsx

Empirical Expectations	Context	Normative Expectations: Low N	Low Excl.	High N	High Excl.
Low	Alarm	72	2	62	2
	Cookbook	63	7	72	11
	Location	67	2	57	1
High	Alarm	54	7	62	10
	Cookbook	61	10	60	9
	Location	55	0	61	7

Table 1: Number of participants in each condition. *N* denotes the number of participants in each condition, and *Excl.* indicates ineligible participants excluded from each condition. Each participant assigned at random to one of twelve conditions.

- *perceived normative expectations*: "Please estimate the percentage of fellow residents in your area who believe it is OK for people to [use their mobile device as their wake-up alarm | use an app on their mobile device as a digital cookbook | allow their mobile device to track their location]." (0-100 in increments of 10)

Normative expectations are evaluative, and are a person's beliefs about what others believe about what behaviors are acceptable or unacceptable. At face value, it may seem strange to think that normative expectations may exist for different uses of one's mobile device, and even stranger to measure this by asking about the prevalence of people who "believe it is OK" to use their mobile device in a particular way. However, people often comply with norms without being consciously aware they are doing so [1]. And, if the research were about a behavior for which it may be more intuitively obvious that social expectations are important, measuring them in this way might seem more straightforward; e.g., "Please estimate the percentage of fellow residents in your area who believe it is OK for people to smoke cigarettes indoors in public places." This study uses similar phrasing and sentence structure to find out whether norms are at work regarding uses of mobile devices that can generate data with privacy implications.

Also, note that the above questions about participants' personal behavior and beliefs related to the technology use context were asked before the vignette was presented. Asking these questions before presenting the vignette ensures that the responses to questions about participants' current beliefs and behaviors were not affected by the experiment manipulation.

The vignette was presented next, followed by a question about the behavior of the main character in the vignette given the situation described:

- *compliance likelihood*: "How likely is this person to [use their mobile device as their wake-up alarm | use an app on their mobile device as a digital cookbook | allow their mobile device to track their location] in this situation?" (Extremely Unlikely (0) - Extremely Likely (10) in increments of 1)

Two additional questions were asked as follow-ups to the vignette, about the believability of the inference presented in the vignette, and the participant's assessment of whether they would use their mobile device as described in the vignette assuming the inference were possible:

- *believability*: "Please indicate your level of agreement with the statement below. If a person [uses their mobile device as their wake-up alarm | uses an app on their mobile device as a digital cookbook | allows their mobile device to track their location], I believe it is possible for the system to [detect how often they snooze or sleep through the alarm | analyze the foods they like to eat and determine how healthy they are | detect how often they visit the restroom]." (Strongly disagree (1) - Strongly agree (5) in increments of 1)

- *personal use likelihood*: "How likely would you be to [use your mobile device as your wake-up alarm | use an app on your mobile device as a digital cookbook | allow your mobile device to track your location], assuming it is possible for the system to [detect how often you snooze or sleep through your alarm | analyze the foods you like to eat and determine how healthy you are | detect how often you visit the restroom]?" (Extremely Unlikely (0) - Extremely Likely (10) in increments of 1)

The final section of the survey asked questions about participants' concern about data privacy using questions from the collection and use subscales of the Concern for Information Privacy Scale (CFIP) [29], past negative privacy/security related experiences, and internet literacy using items based on a measure by Hargittai [9], as well as other questions not used in this paper. The full survey instrument is available online, with the replication materials for the experiment.

4 Results

4.1 Self-Reported Behavior and Beliefs in Each Context

A majority of participants in both the alarm (177 of 250) and location (148 of 240) conditions answered Yes to the personal behavior question, indicating that they either use their mobile device as a wake-up alarm or allow it to track their location. Most participants across all three contexts also reported that they believe it is OK for people to use their mobile devices for these things. This pattern was most pronounced for the alarm conditions, where 71% of participants said they used their mobile device for this, and 98% said it is OK for others to do so. In contrast, 96% of participants in the cookbook conditions said it is OK to use their mobile device as a digital cookbook,

	Personal Behavior		Personal Beliefs	
	No	*Yes*	*No*	*Yes*
Alarm	29%	71%	2%	98%
Cookbook	75%	25%	4%	96%
Location	38%	62%	28%	72%

Table 2: Percent of participants in each context condition who reported doing the behavior (Personal Behavior) and believing it is OK for others to do the behavior (Personal Beliefs).

	Empirical Expect.			Normative Expect.		
Context	*Mean*	*Med.*	*SD*	*Mean*	*Med.*	*SD*
Alarm	56.9	60	27.3	68.6	80	26.9
Cookbook	32.7	30	23.4	61.5	60	30.9
Location	52	50	22.1	50	50	23.2

Table 3: Descriptive statistics for participants' empirical expectations (beliefs about what others do) and normative expectations (beliefs about what others believe) in each context condition, on a scale from 0 to 100 in increments of 10.

but only 25% of participants said they actually did. And while 72% of participants in the location conditions said it is OK to allow one's mobile device to track one's location, ten percent fewer (62%) said they personally allowed this. These results show that overall, most participants were comfortable with these contexts for using mobile devices, and saw nothing wrong with others using their mobile devices in these ways. Table 2 presents the percent of participants who answered Yes versus No to the questions about personal behavior and beliefs. Replication materials, including the data and code to reproduce the analyses, are available online.

Participants were also asked to estimate the how common the behavior is among other people in their area (perceived empirical expectation), and to what extent people believe that it is OK to do the behavior (perceived normative expectation). For example, participants who saw vignettes about the location context were asked to estimate the percentage of others in their area who allow their mobile device to track their location, and who believe that it is OK to allow this. In the alarm and cookbook contexts, the mean estimated percentage was higher for perceived normative expectations (alarm: 68.6, cookbook: 61.5) than for perceived empirical expectations (alarm: 56.9, cookbook: 32.7). In other words, participants believed that it is more common for people to approve of these uses of mobile devices than to actually engage in using them in these ways. But in the location context, the mean empirical and normative expectations were about the same (empirical: 52, normative: 50). These results show that participants in each context believed there are some social expectations associated with the behaviors; but, they believed fewer people believe location tracking is acceptable than the alarm and cookbook contexts. Table 3 shows the average estimated percentages for participants' empirical and normative expectations in each context.

4.2 Privacy-related choices are correlated with beliefs about others' behavior (RQ1)

Participants' answers to the questions about their personal beliefs and their perceived empirical and normative expectations can be used to identify correlations between these factors and their self-reported behavior. This allows us to investigate whether social expectations are associated with participants' own choices to use their mobile devices as wake-up alarms, digital cookbooks, or to allow location tracking.

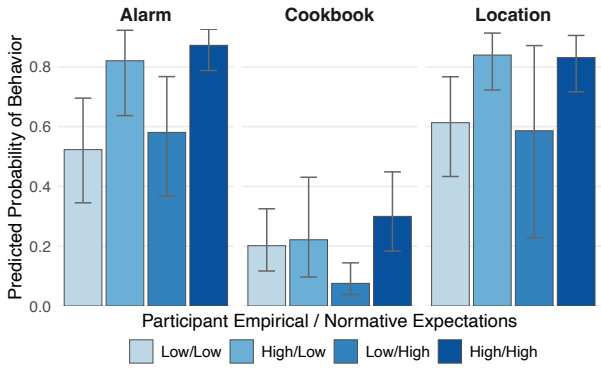


Figure 1: Predicted probabilities from the logistic regression model showing that perceived empirical and normative expectations are associated with a greater likelihood of complying with the behavior in the context condition participants were assigned to (RQ1). Error bars represent 95% confidence intervals.

The perceived empirical and normative expectations variables were each split into high vs. low categories at the median of each variable, so the results of this analysis would be more comparable with the results of the experimentally manipulated empirical and normative expectations analyses presented in the next section. The 2 x 2 high vs. low empirical and normative expectations variables were then recoded into a single categorical variable with four levels for use in the regression model (high empirical/high normative, high/low, low/high, low/low). Note that it is incorrect to assume that low perceived normative expectations (a low estimated percentage of others who believe the behavior is OK) means that a high proportion believe it is not OK. It could be that reporting a low percentage means that participants are not knowledgeable about others' beliefs, or that no norm exists.

A logistic regression model was used to identify factors associated with self-reported use, which was the dependent variable. The model has a categorical predictor for the context (alarm, cookbook or location) and a categorical predictor

for the combination of perceived empirical and normative expectations. It also includes the interaction between context and expectations, and controls for personal beliefs about whether the behavior is OK, demographic variables gender and age, the number of negative security/privacy experiences the participant reported, privacy concern, and internet literacy. If there is a relationship between perceived empirical and/or normative expectations and the dependent variable, then we can conclude that social expectations exist and may influence whether participants use their mobile devices as the study contexts described.

The intercept in the model represents the category combination of alarm (context), low/low (perceived empirical / normative expectations), no (personal belief), and man (gender). Positive coefficients in the model indicate greater odds that the participant would self-report doing the behavior, as compared to the intercept. The model results indicate that believing the behavior is OK has a strong, positive influence on the odds that the participant will report that they do the behavior. The odds that participants would report doing the behavior were 14.6 times higher (*coef* = 2.67) when they believe that it is OK to do the behavior.

However, even with that predictor in the model, a high perceived empirical expectation was associated with a greater odds of doing the behavior. The coefficient of 1.43 for high empirical/low normative expectations indicates that the odds are 4.2 times higher that participants report doing the behavior when they perceive that a high percentage of others do the behavior, even if the perceived normative expectations (belief that others approve of the behavior) are low. When participants perceive that both empirical and normative expectations are high, the odds of reporting the behavior are 6.2 times higher (*coef* = 1.82).

Also of note are the negative coefficients for the cookbook context, and the interaction between the cookbook context and the perceived empirical/normative expectations. While most of the interaction coefficients are not statistically significant, these coefficients do illustrate that participants were in general less likely to report using their mobile device as a cookbook and also perceiving that others do so. This is reflected in the model as lower odds of doing the behavior in the cookbook context even when the empirical expectation is high, in contrast to the the other two contexts. Table 4 presents the full regression results for this model, in the leftmost column.

These results show that participants' empirical expectations—their beliefs about how others use their mobile devices—are related to whether they use their mobile devices in similar ways. However, participants' own normative expectations—beliefs about whether others believe it is OK to do the behavior—did not have any effect beyond the effect of empirical expectations. This can clearly be seen in Figure 1, which presents the predicted probabilities from the model for men with personal beliefs that it is OK to do the behavior. Where empirical expectations are low, the likelihood of doing the behavior is lower than where empirical expectations are high. Normative expectations are not associated with greater likelihood beyond empirical expectations. This means that seeing others around them doing the behavior—using their mobile devices as alarms, cookbooks, or allowing location tracking—is strongly associated with the participants doing the behavior themselves.

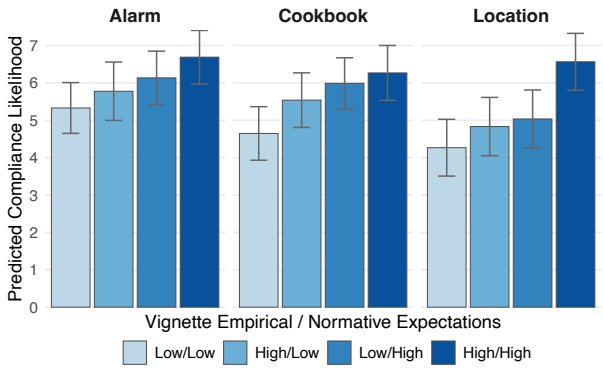


Figure 2: Predicted values from the OLS model showing that experimentally manipulated social expectations increase compliance likelihood (RQ2). Error bars represent 95% confidence intervals.

4.3 Norms support conforming with others' privacy-related choices (RQ2)

The previous model showed that social expectations are probably a factor in participants' choices to use their mobile phones for purposes that may allow sensitive data to be collected about them. However, participants were not asked about their awareness of the possibility of such inferences, and the previous model can only identify correlations between the predictors and the dependent variable. In order to determine whether social expectations are "causally relevant" [2] for participants' behavior, empirical and normative expectations were experimentally manipulated in the vignettes. Each vignette also presented a possible inference that could be made as a result of using one's mobile device as the vignette described. If participants reported a higher likelihood that the main character in the vignette would do the behavior when empirical and/or normative expectations in the vignette are high than when they are low, then we can conclude that social expectations affect the likelihood of compliance with the behavior.

An OLS regression model was used to find out if a causal relationship exists between the experimentally manipulated empirical and normative expectations presented in the vignette and the likelihood that the main character in the vignette would do the behavior. The dependent variable, compliance likelihood, was on an 11 point scale from Extremely Unlikely

	RQ1: Personal behavior (*logistic*)	RQ2: Compliance likelihood (*OLS*)	RQ3: Personal use likelihood (*OLS*)
Context: cookbook	-1.468** (0.479)	-0.681 (0.470)	-1.921*** (0.494)
Context: location	0.369 (0.495)	-1.062* (0.483)	-2.670*** (0.507)
Expectations: high empirical/low normative	1.428* (0.584)	0.448 (0.494)	-1.066* (0.519)
Expectations: low empirical/high normative	0.234 (0.548)	0.805• (0.471)	-0.907• (0.495)
Expectations: high empirical/high normative	1.828*** (0.457)	1.359** (0.474)	-0.606 (0.498)
Personal beliefs: Yes	2.679*** (0.402)	1.298*** (0.354)	2.313*** (0.372)
Gender: woman	0.212 (0.232)	0.241 (0.236)	0.735** (0.248)
Age	-0.043*** (0.008)	-0.024*** (0.007)	-0.031*** (0.008)
Num negative security/privacy experiences	0.216** (0.074)	-0.093 (0.075)	0.130• (0.079)
Privacy concern	-0.269 (0.169)	0.051 (0.166)	-0.362* (0.175)
Internet literacy	0.412** (0.137)	0.232• (0.133)	0.646*** (0.140)
Believability		0.492*** (0.093)	0.683*** (0.098)
cookbook * high empirical/low normative	-1.308 (0.809)	0.443 (0.695)	1.120 (0.730)
location * high empirical/low normative	-0.233 (0.740)	0.118 (0.706)	1.451• (0.742)
cookbook * low empirical/high normative	-1.361• (0.708)	0.535 (0.665)	0.959 (0.698)
location * low empirical/high normative	-0.346 (1.014)	-0.036 (0.684)	0.977 (0.719)
cookbook * high empirical/high normative	-1.301* (0.607)	0.260 (0.681)	1.128 (0.715)
location * high empirical/high normative	-0.696 (0.644)	0.938 (0.676)	0.716 (0.710)
Intercept	-0.731 (0.939)	2.866** (0.984)	3.195** (1.033)
Observations	739	739	739
R^2	0.38 (McFadden's)	0.194	0.333

• p<0.1; * p<0.1; ** p<0.05; *** p<0.01

Table 4: Regression coefficients (and standard errors) for the three regression models. RQ1 focuses on participants' current behavior, RQ2 on their estimate of compliance in the vignette situation, and RQ3 on their behavioral intentions given the information about the inference in the vignette. For RQ1, the empirical and normative expectations are the participant's self-report; for RQ2 and RQ3 they are experimentally manipulated via the vignette. Seven observations with gender category "Not Reported" were excluded from all regressions; these observations were evenly spread across the experiment conditions due to random assignment.

(0) to Extremely Likely (10) in increments of 1 (M = 5.6, *Median* = 6, SD = 3). This model has the same predictors as the previous model, except the empirical and normative expectations experimentally manipulated in the vignette are used instead of the self-reported perceived empirical and normative expectations. This model also has an additional predictor: believability, which represents the participant's evaluation of whether they believe that the inference in the vignette is possible (M = 3.4, *Median* = 4, SD = 1.2). Believability was lower for the location condition (M = 2.90) than the alarm (M = 3.63) or cookbook (M = 3.64) conditions. Table 4 presents the results of this model in the middle column.

High empirical and normative expectations presented in the vignette both caused an increase in compliance likelihood in the experiment. All social expectations categories (high/low, low/high, and high/high) had positive coefficients, indicating an increase in compliance likelihood when compared with the reference category of low/low. The coefficient was smallest and not statistically significant where normative expectations were low (*coef* = 0.44). However, in the low/high category, compliance likelihood was 0.80 points higher than in the low/low condition, and 1.35 points higher when both experimentally manipulated empirical and normative expectations were high (high/high). This indicates that a causal relationship exists between both types of social expectations and compliance likelihood in the experiment.

Like the previous model, the influence of the participant's personal belief that it is OK to use one's mobile device as described in the vignette had a strong, positive influence on compliance likelihood, which was 1.29 points higher when personal belief was Yes than when it was No. Believability was also important: compliance likelihood was 0.49 points higher for each 1-point increase in believability.

Because context was randomly assigned to participant, we can also draw causal conclusions about the impact of the use context on compliance likelihood. The coefficients for both the cookbook context and the location context are negative, indicating that compliance likelihood was lower than in the baseline alarm context. For the location context, the coefficient was large and statistically significant, indicating that compliance likelihood is 1.06 points lower for location vignettes than alarm vignettes. None of the coefficients for the interaction between context and empirical/normative expectations were statistically significant.

Figure 2 shows the pattern of these results in the form of predicted values calculated from the model. In all three context conditions, compliance likelihood is highest where both empirical and normative expectations are high, and lowest

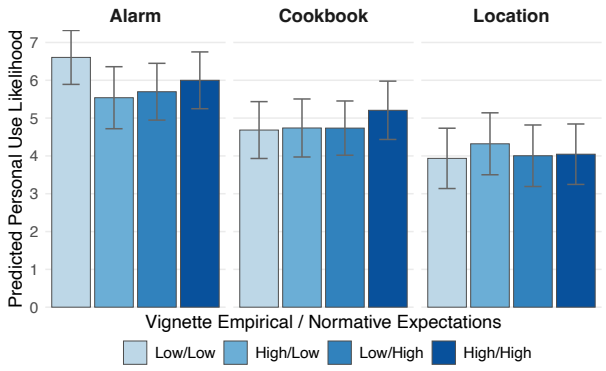


Figure 3: Predicted values from the OLS model showing that participants' estimate of how likely they would be to use their mobile device as described in the vignette decreases as the inferences in the vignette become more invasive and potentially harmful (RQ3). Error bars represent 95% confidence intervals.

where both are low, after controlling for the other variables in the model, including the participant's personal beliefs. The vignettes asked participants to imagine what someone like them would do, assuming the situation in the vignette were true. Compliance likelihood can therefore be interpreted as a measure of what the participants themselves would do [1]. This means that this model provides evidence that not only are social expectations related to participants' behavior, they are causally related. In other words, believing that others use technologies in ways that allow invasive data collection and also approve of doing so increases the likelihood that an individual will also allow this themselves. This provides further evidence that norms exist supporting the use of apps on mobile devices that make potentially unwanted, invasive inferences.

4.4 More invasive inferences are associated with lower behavioral intentions (RQ3)

A final question remains about whether the inferences in the vignettes are really unwanted. Participants in the study showed overall concern regarding data privacy, as measured by items from the collection and use subscales of the Concern for Information Privacy survey instrument (CFIP). The overall mean across all of the questions asked in this experiment was 4.3 out of 5 (higher means more concerned), and there were no differences across the context conditions. But, this does not necessarily mean an objection to the specific inferences mentioned in this study. See the replication materials, available online, for descriptive statistics about participants' responses to the CFIP questions.

In addition to measuring the compliance likelihood of the main character in the vignette, the survey also asked participants to estimate how likely they themselves would be to do the behavior described in the vignette, assuming the inferences were actually possible. This question was essentially about participants' behavioral intentions, measured after an awareness intervention (the vignette) informing them about possible inferences. If participants were opposed to the inferences in the vignette they read, this would be reflected in their measured behavioral intentions.

Context	M	Median	SD
alarm	6.4	7	3.3
cookbook	5.4	5.5	3.1
location	3.5	3	3.2

Table 5: Personal use likelihood descriptive statistics.

An OLS regression model was used to identify whether a relationship exists between the experimentally manipulated empirical and normative expectations and participants' self-report of how likely they would be to use their mobile devices in the way describes in the vignette, (personal use likelihood) if the inferences presented in the vignettes were possible. The model has personal use likelihood as the dependent variable, which used the same response category structure as the compliance likelihood measure in the previous section. The predictors in the model are also identical to the compliance likelihood OLS model. Table 4 presents the results of this model in the rightmost column.

The inference in the alarm context (the baseline context condition) focused on tracking oversleeping, which was expected to be the least concerning inference to participants based on Rader's interview study [20]. In the cookbook context, the inference was about how healthy the participant is based on the foods in the recipes they cook, which could be more concerning but also potentially helpful to someone who wants to adopt healthier eating habits. The inference in the location condition about detecting bathroom behavior was expected to be fairly unacceptable to participants, because it was unacceptable to most of Rader's interview participants. The coefficients in the model for the cookbook context (*coef* = -1.92) and location context (*coef* = -2.67) are both negative, large, and statistically significant, showing the expected pattern.

The means for the personal use likelihood variable, presented in Table 5, clearly illustrate this relationship. On average, personal use likelihood was highest for the alarm context (M = 6.4), lower for the cookbook context (M = 5.4), and lowest in the location context (M = 3.5). These results are an indication that where the inferences are more invasive, participants reported that they would be less likely to behave in ways that would enable the inferences to be made.

In contrast to the other two models, the coefficients for the experimentally manipulated empirical and normative expectations conditions are negative. Only the coefficient for high

empirical / low normative expectations is statistically significant, but it is fairly large (*coef* = -1.06). These results present an inconsistent pattern: if social expectations (empirical or normative) were consistently influential, then this would be apparent in the coefficient for high empirical / high normative as well. In addition, none of the coefficients for the interaction between context and social expectations are statistically significant. The only conclusion that can be drawn from this model is that the social expectations presented in the vignettes do not have a clear relationship with participants' behavioral intentions related to technologies that generate potentially invasive inferences.

As in the previous OLS model for RQ2, believability of the inference had a positive, statistically significant effect on personal use likelihood (*coef* = 0.68). Rader [20] wrote that when inferences were thought to be useful, for example for helping users to correct bad habits or improve their health, the reaction to the inferences was more positive. This could be an indication that participants who believed the inferences were possible may have been more enthusiastic about potential benefits from the inferences. Finally, in this model, like the other two, personal expectation (whether the participant believes it is OK to use their mobile device as an alarm, a cookbook, or to track their location) had a strong, positive effect.

Figure 3 clearly shows that the highest predicted use likelihood is for the alarm conditions, followed by the cookbook conditions and then the location conditions. This is different from the results of the RQ1 logistic model (see Figure 1), where the predicted probability of using one's mobile device as a cookbook was much lower than the other two context conditions. Bicchieri et. al [2] argue that using hypothetical scenarios about a protagonist that is not the participant but is similar to them frees participants from considering the details of their own lives and situations when considering how the person in the vignette would react given the described situation. It could be that participants considered other contextual factors beyond the experimentally manipulated social expectations in the vignette when answering about their own behavioral intentions. These may reflect a positive reaction to the idea of a digital cookbook, but participants may lack actual opportunities to use their devices in this way. It is not surprising that participants in the location condition would be the least comfortable with the associated inference, which was about detecting bathroom visits. Overall, this model provides evidence that the inference awareness intervention presented in the vignette was associated with lower participant willingness to do the behavior where the inferences were more invasive.

4.5 Limitations

This research has several limitations. First, because this is a survey-based experiment, participants' answers to the questions are self-report and reflect their beliefs and their perceptions of their own behavior, but should not be interpreted as direct evidence of their actual or future behaviors.

Second, sampling choices limit the generalizability of the results, in a couple of ways. Participation was limited to people who reported that they did not have high-tech related expertise, because normative influences may be more important factors for non-experts in their choices to use certain technologies or allow data collection. People with training or work experience in a high-tech related field may react differently to the vignettes due to knowing more about how inferences are generated. Also, the experiment used two recruiting quotas, age and gender. This means that the sample is not statistically representative of the internet-using population of the United States for other demographic characteristics like ethnicity, income, education, etc. While this sample has more external validity than a sample where participants were selected entirely based on convenience, this study did not use probability sampling to select participants. Therefore, results should not be generalized to experts, or used to make claims about the broader U.S. population.

In addition, while the selection of the three vignette contexts was based on prior research [20], the specifics of the vignettes participants were asked to react to in the experiment undoubtedly had an influence on the results. It is possible that social influence on data privacy decisions varies by context, and if different contexts had been selected for the experiment the results might have been different.

Finally, the three vignette contexts (alarm, cookbook, location) are different from each other in a number of ways. For example, the alarm and cookbook contexts include a purpose for the behavior (wake-up alarm and digital cookbook), while the location context focuses on a type of information / data (location) without specifying a purpose. The three inferences across the vignettes also vary, in that information about oversleeping may be seen as less of a privacy violation than information about bathroom behavior. This means that while the experiment design supports an interpretation that contextual factors influence privacy-related choices, it is not possible to determine from this experiment which specific contextual factors are responsible for the effect.

5 Discussion

Pluralistic ignorance occurs where individuals engage in a behavior they dislike because of social expectations that they do so. This study investigated whether pluralistic ignorance may be occurring with respect to data privacy and the use of mobile devices. Under conditions of pluralistic ignorance, incorrect beliefs that others approve of the use of potentially invasive technologies would conflict with individuals' privacy preferences and concern. Complying with the social pressure and adhering to the norm would mean behaving contrary to one's own beliefs about privacy.

The results of this study show that a positive correlation exists between using one's mobile device in ways that could compromise one's privacy, and a belief that others use their mobile devices in similar ways. This is one aspect of pluralistic ignorance: believing that a behavior one engages in is commonly done by others too. In addition, the study found evidence of a causal relationship between both empirical and normative expectations and the likelihood of complying with the norm and engaging in the behavior described in the hypothetical vignette. Normative expectations are beliefs about what others believe should be done. The fact that these beliefs were causally related with compliance likelihood in the experiment means that the social influence is more than just imitation—compliance is also affected by the approval/disapproval of others. This is another aspect of pluralistic ignorance: believing that others approve of the behavior.

The third aspect of pluralistic ignorance is a private dislike of the behavior. Participants in the experiment expressed general privacy concern, and also said they would be less likely to use their mobile device as the vignette described where the inferences in the vignette were more invasive and potentially harmful. Taken together, these results suggest that pluralistic ignorance is taking place with respect to data privacy.

A common approach to intervening to help end users make privacy choices that are more consistent with their preferences is providing information to increase awareness of data collection and inferences. This approach is based on the idea that data privacy is an individual decision about control, and education or literacy campaigns can change the inputs to those decisions as a way to change the outcome of the decisions. Many informational campaigns seek to change attitudes towards a behavior by providing previously unknown information about the dangers or harms of that behavior (e.g., alcohol abuse campaigns that focus on persuasion by conveying information about the dangers of alcohol [7,27]). However, the results of this study show that while data privacy results from individual choices, there are social influences on those choices. This means that data privacy is also social, and interventions are needed that take this into account.

Recent research has suggested that peer information sharing might help people make better privacy decisions [5,13,15,16]. Unfortunately, in a pluralistic ignorance situation, sharing information about others' behavior is likely to backfire, because pluralistic ignorance is inherently self-reinforcing. If everyone uses invasive technologies despite privately disliking doing so, sharing information about others' behavior would reinforce the very norm which influences people to behave contrary to their true preferences. In other words, more information about what others do would actually perpetuate the problem by strengthening the norm. Under conditions of pluralistic ignorance, more transparency about others' true beliefs, not their behavior, is needed. However, many approaches to social cybersecurity and privacy do indeed focus on helping individuals via more transparency about others' behavior.

Changing norms under conditions of pluralistic ignorance is hard, because nobody wants to be the first one to express their true preferences and behave contrary to the norm [1]. This prevents people from realizing they are not alone in their dislike of the norm; and it also prevents collective action towards a solution. People cannot coordinate if they all believe they are the only one who thinks they way they do [12]. To encourage people to make different choices, it is necessary to counteract each person's incorrect assumption that everyone approves of the invasive data collection but them.

The most common approach to changing the norm under these conditions is to expose the pluralistic ignorance: people need to know they are not alone, and that others also disapprove and want to protect their privacy and their data. Previous research has found some success in informational campaigns focused not on reasons to adopt the behavior change, but on the true beliefs of others [27]. The goal of this type of intervention is to correct the misperception that each individual is the only one who dislikes the behavior. In addition to informational campaigns, "trendsetters" can also be successful, but only if the conditions are right. A successful trendsetter for counteracting pluralistic ignorance must be an independent thinker, not sensitive to being judged by others, and believe that going against the norm will do some good. They also need to be well positioned in their social network to reach enough people such that when they go against the norm, it makes deviant behavior seem less risky for enough people that the norm falls apart [3].

And here is the final challenge: protecting one's privacy is by nature an action that is not very visible. As a way of combatting pluralistic ignorance, mechanisms must be developed to make protecting privacy more visible without compromising it. For example, the Facebook "I Voted" button was reported to have significantly increased voter turnout in the U.S. in 2010 even in light of voter apathy [4]. Voting is ostensibly a behavior that is private and not observed, potentially making this instance an analogy to interventions focused on expressing one's true beliefs about data privacy. Providing a mechanism that would allow visibility into people's desire to protect their information would also require efforts to put a positive spin choices to protect one's information, to avoid assumptions based on the 'nothing to hide' myth [31] that only bad people have reasons to want to protect their privacy.

Clearly, there are other barriers to improving data privacy than pluralistic ignorance. As much literature has discussed, people do not have a lot of options for truly protecting their data, especially in workplace and education settings. And, notice and choice ensures that people are asked for consent before they have a good idea of what the privacy implications of their consent might be. But, solutions to these issues will arguably be less successful and may even fail if they ignore the role that social expectations play in data privacy.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. CNS-1524296.

References

[1] Cristina Bicchieri. *Norms in the Wild*. Oxford University Press, 2016.

[2] Cristina Bicchieri, Enrique Fatas, Abraham Aldama, Andrés Casas, Ishwari Deshpande, Mariagiulia Lauro, Cristina Parilli, Max Spohn, Paula Pereira, and Ruiling Wen. In science we (should) trust: Expectations and compliance across nine countries during the COVID-19 pandemic. *PLOS ONE*, 16(6):e0252892, 2021.

[3] Cristina Bicchieri and Alexander Funcke. Norm Change: Trendsetters and Social Structure. *Social Research: An International Quarterly*, 85(1):1 – 22, 09 2018.

[4] Robert M Bond, Christopher J Fariss, Jason J Jones, Adam D I Kramer, Cameron Marlow, Jaime E Settle, and James H Fowler. A 61-million-person experiment in social influence and political mobilization. *Nature*, 489(7415):295 – 298, 2012.

[5] Chhaya Chouhan, Christy M LaPerriere, Zaina Aljallad, Jess Kropczynski, Heather Lipford, and Pamela J Wisniewski. Co-designing for Community Oversight: Helping People Make Privacy and Security Decisions Together. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–31, 2019.

[6] Robert B Cialdini, Linda J Demaine, Brad J Sagarin, Daniel W Barrett, Kelton Rhoads, and Patricia L Winter. Managing social norms for persuasive impact. *Social Influence*, 1(1):3–15, March 2006.

[7] Dale T. Miller Deborah A. Prentice. Pluralistic ignorance and the perpetuation of social norms by unwitting actors. *Advances in Experimental Social Psychology*, 28:161–209, 1996.

[8] Shirley Gaw, Edward W. Felten, and Patricia Fernandez-Kelly. Secrecy, flagging, and paranoia: Adoption criteria in encrypted email. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 591–600, 2006.

[9] E Hargittai. An Update on Survey Measures of Web-Oriented Digital Literacy. *Social Science Computer Review*, 27(1):130 – 137, 2008.

[10] Samantha Hautea, Anjali Munasinghe, and Emilee Rader. That's not me: Surprising algorithmic inferences. In *Poster presented at the 2020 Symposium on Usable Privacy and Security*, 2020.

[11] Kyle Irwin and Brent Simpson. Do Descriptive Norms Solve Social Dilemmas? Conformity and Contributions in Collective Action Groups. *Social Forces*, 91(3):1057–1084, February 2013.

[12] Esther Michelsen Kjeldahl and Vincent F Hendricks. The sense of social influence: pluralistic ignorance in climate change. *EMBO Reports*, 19:e47185, 2018.

[13] Isadora Krsek, Kimi Wenzel, Sauvik Das, Jason I Hong, and Laura Dabbish. To Self-Persuade or be Persuaded: Examining Interventions for Users' Privacy Setting Selection. *CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2022.

[14] Dale T. Miller and Cathy McFarland. Pluralistic Ignorance: When Similarity is Interpreted as Dissimilarity. *Journal of Personality and Social Psychology*, 53(2):298–305, 1987.

[15] Pardis Emami Naeini, Martin Degeling, Lujo Bauer, Richard Chow, Lorrie Faith Cranor, Mohammad Reza Haghighat, and Heather Patterson. The Influence of Friends and Experts on Privacy Decision Making in IoT Scenarios. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1 – 26, 11 2018.

[16] Bettina Nissen, Victoria Neumann, Mateusz Mikusz, Rory Gianni, Sarah Clinch, Chris Speed, and Nigel Davies. Should I Agree? Delegating Consent Decisions Beyond the Individual. Proceedings of the CHI Conference on Human Factors in Computing Systems, pages 1 – 13, 2019.

[17] Helen Nissenbaum. Privacy as Contextual Integrity. *Washington Law Review*, 79:119–158, 2004.

[18] Sandra Petronio. *Boundaries of Privacy: Dialectics of Disclosure*. State University of New York Press, Albany, NY, 2002.

[19] President's Council of Advisors on Science and Technology. Big data and privacy: a technological perspective. Technical report, May 2014.

[20] Emilee Rader. Normative and Non-Social beliefs about sensor data: Implications for collective privacy management. In *Eighteenth Symposium on Usable Privacy and Security (SOUPS 2022)*, Boston, MA, August 2022. USENIX Association.

[21] Emilee Rader, Samantha Hautea, and Anjali Munasinghe. I have a narrow thought process: Constraints on explanations connecting inferences and self-perceptions. In *Sixteenth Symposium on Usable Privacy and Security (SOUPS 2020)*. USENIX Association, 2020.

[22] Nichola J. Raihani and Vaughan Bell. An evolutionary perspective on paranoia. *Nature Human Behaviour*, 3(2):114–121, 2019.

[23] Priscilla M Regan. *Legislating Privacy: Technology, Social Values, and Public Policy*. The University of North Carolina Press, Chapel Hill, NC, 1995.

[24] Priscilla M Regan. Privacy as a Common Good in the Digital World. *Information, Communication & Society*, 5(3):382–405, 2002.

[25] Neil M. Richards. Four privacy myths. In Austin Sarat, editor, *A World Without Privacy: What Law Can and Should Do?*, pages 33–82. Cambridge University Press, 2015.

[26] Rikki H. Sargent and Leonard S. Newman. Pluralistic ignorance research in psychology: A scoping review of topic and method variation and directions for future research. *Review of General Psychology*, 25(2):163–184, 2021.

[27] Christine M. Schroeder and Deborah A. Prentice. Exposing pluralistic ignorance to reduce alcohol use among college students. *Journal of Applied Social Psychology*, 28(23):2150–2180, 1998.

[28] Jacob Shamir and Michal Shamir. Pluralistic Ignorance Across Issues and Over Time: Information Cues and Biases. *The Public Opinion Quarterly*, 61(2):227–260, 1997.

[29] H. Jeff Smith, Sandra J. Milberg, and Sandra J. Burke. Information privacy: Measuring individuals' concerns about organizational practices. *MIS Quarterly*, 20(2):167–196, 1996.

[30] Daniel Solove. The myth of the privacy paradox. *Geo. Wash. L. Rev.*, 89(1), 2021.

[31] Daniel J Solove. "i've got nothing to hide" and other misunderstandings of privacy. *San Diego L. Rev.*, 44:745, 2007.

[32] Daniel J Solove. Introduction: Privacy self-management and the consent dilemma. *126 Harvard Law Review*, pages 1880–1903, 2013.

[33] Justin Wu and Daniel Zappala. When is a Tree Really a Truck? Exploring Mental Models of Encryption. In *Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018)*, pages 1–16, 2018.

Appendix

	N	%		*N*	%
Age			**Employment**		
18–20	36	5%	Employed full time	243	33%
21–44	305	41%	Employed part time	112	15%
45–64	232	31%	Unemployed or Disabled	182	24%
65+	173	23%	Unemployed or Disabled	182	24%
Gender			Retired	179	24%
Man	361	48%	Student	30	4%
Woman	378	51%	**Income (USD)**		
No Gender Reported	7	1%	Less than $25,000	167	22%
Education			$25,000 to $34,999	122	16%
Some High School	22	3%	$35,000 to $49,999	115	15%
High School Grad	500	67%	$50,000 to $74,999	155	21%
College Grad	152	20%	$75,000 to $99,999	89	12%
Postgraduate degree	72	10%	$100,000 to $149,999	75	10%
Ethnicity			$150,000 to $199,999	13	2%
White	575	77%	$200,000 or more	10	1%
Black or African American	75	10%	**Residential Area**		
Multiple Ethnicities	31	4%	Village or Countryside	113	15%
Hispanic, Latino or Spanish	29	4%	Small or Mid-Size Town	418	56%
Asian or Pacific Islander	26	4%	Large City	215	29%
Native American Alaskan	5	1%			
Other or Not Specified	5	1%			

Table 1: Participant demographics.

Distrust of big tech and a desire for privacy: Understanding the motivations of people who have voluntarily adopted secure email

Warda Usman
Brigham Young University

Jackie Hu
Brigham Young University

McKynlee Wilson
Brigham Young University

Daniel Zappala
Brigham Young University

Abstract

Secure email systems that use end-to-end encryption are the best method we have for ensuring user privacy and security in email communication. However, the adoption of secure email remains low, with previous studies suggesting mainly that secure email is too complex or inconvenient to use. However, the perspectives of those who have, in fact, chosen to use an encrypted email system are largely overlooked. To understand these perspectives, we conducted a semi-structured interview study that aims to provide a comprehensive understanding of the mindsets underlying adoption and use of secure email services. Our participants come from a variety of countries and vary in the amount of time they have been using secure email, how often they use it, and whether they use it as their primary account. Our results uncover that a defining reason for adopting a secure email system is to avoid surveillance from big tech companies. However, regardless of the complexity and accuracy of a person's mental model, our participants rarely send and receive encrypted emails, thus not making full use of the privacy they could obtain. These findings indicate that secure email systems could potentially find greater adoption by appealing to their privacy advantages, but privacy gains will be limited until a critical mass are able to join these systems and easily send encrypted emails to each other.

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.
USENIX Symposium on Usable Privacy and Security (SOUPS) 2023.
August 6–8, 2023, Anaheim, CA, USA

1 Introduction

There are over 319 billion[1] emails sent every day. These emails are transmitted and stored primarily in plaintext, and are therefore subject to a wide variety of threats, including surveillance, modification, commercial analysis, and theft. Emails sent and received by larger providers are often encrypted when they are sent between email servers, but this is not universally deployed, can be circumvented, and still leaves emails vulnerable to attacks where they are stored [9]. This insecure communication creates a variety of security and privacy threats for users.

To protect the privacy and security of messages, experts have been suggesting end-to-end encryption (E2EE) and encrypted storage for decades now. Despite these efforts, the use of E2EE for email has remained relatively scarce [45].

The research community has generally focused on improving the usability of secure email systems, believing that this was the primary obstacle to adoption. This work began with a seminal paper by Whitten and Tyger [49], showing that users made mistakes when interacting with key pairs. These problems continued to plague systems based on PGP for years [36, 40], but recent work has shown how to provide usable secure email systems by automating user interactions with keys and certificates as much as possible [14,37,38]. Current web-based systems, such as Proton Mail and Tutanota, utilize automation and have interfaces that are largely similar to popular email sites like Gmail, so usability is unlikely to remain a significant obstacle to adoption.

A growing body of work has demonstrated that a variety of factors beyond usability affect adoption of secure email. In a broad look at secure communication tools [1], the primary obstacles were found to be fragmented user bases, lack of interoperability, and low quality of service. Lack of advertising is also an issue; a large number of people are still unaware of the existence of any secure email services [45]. Further, users

[1] From the Email Statics Report, 2021-2025, by The Radicati Group, https://www.radicati.com/wp/wp-content/uploads/2020/12/Email-Statistics-Report-2021-2025-Executive-Summary.pdf

resist adopting secure email due to their incomplete threat models, misaligned incentives, and due to lack of understanding about the secure email architecture [34]. Other factors which go beyond an individual user also contribute [4]. For example, secure email requires global interoperability among heterogeneous clients and systems to be useful. Additionally, many stakeholders of secure email do not agree on what properties to provide, which hinders development of a ubiquitous protocol for secure email. Another factor is that encrypted storage of email makes search of encrypted archives and scanning for spam and malware more difficult, which might cause some users to stick with traditional unencrypted methods.

Today millions of people do use secure web-based email systems and some businesses use S/MIME integrated into email clients such as Outlook. This is a significant improvement over past decades, but still well short of the billions using standard email and the billions using secure messaging apps such as WhatsApp. However, many of the lessons learned from the adoption of secure messaging do not provide similar pathways for adoption of secure email systems. First, users with no interest in the security or privacy features of secure messaging apps have primarily adopted one because their regular communication partners used it [1]. This is easier to do with secure messaging applications, since they are walled gardens, which means that users can only communicate with those using the same provider. Secure email, on the other hand, must remain interoperable with a wide variety of non-secure email systems and clients in order to be useful; a friend using secure email doesn't require you to join that same system in order to communicate. Another factor motivating adoption of secure messaging apps is that they enable users to avoid texting fees for international messaging. This also doesn't apply to email since it is generally a free service.

Our goal in this work is to better understand those relatively unusual people who choose to use a secure email service such as Proton Mail or Tutanota. While prior work has focused on the *lack* of adoption, these people have made the choice to use a system offering privacy and security benefits when free, less secure, and less private tools are readily available. Moreover, these users must operate in a world where the vast majority of their emails are likely going to other people who do *not* use a secure system, in contrast to the walled garden offered by a secure messaging app. Talking to users who have made this choice can help us to understand their motivations and provide insight into whether more people could follow their path.

We identified the following research questions:

1. Why do people voluntarily adopt secure email systems?
2. What threat models do people have, meaning their conception of attackers and the harms they can impose, and what steps do they take to mitigate these harms?
3. What mental models do people have of secure email systems and their capabilities? We particularly want to understand perceptions of what security and privacy means within the context of email and how secure email systems provide security and privacy.
4. Do people use the secure email services effectively and what obstacles they encounter in trying to do so?

To answer these questions, we conducted an interview study among users of secure email systems, primarily Proton Mail. We interviewed 25 participants who currently use Proton Mail, from 12 different countries. Our interview focused on answering the four questions listed above, thus discussing their reasons for adoption, their mental models, their threat models, and their usage of secure email. We analyzed the interviews using a mix of inductive and deductive coding, depending on which applied best to a given research question.

Our findings indicate that motivations to adopt a secure email system include a combination of distrust of big tech companies, aversion to targeted advertising, various notions of privacy, affordances, trust in companies that offer privacy, and a desire to align decisions with companies that share their values. Privacy resonates strongly with the participants, with Proton Mail seen as one way they can avoid big tech companies or obtain a particular privacy benefit. Participants recognized that major harms could come from government surveillance or hackers stealing their email, but were motivated by threats they felt were more likely, such as the general surveillance economy. These feelings were consistent both among those who had only a limited understanding of how a secure email system works and those who had accurate, detailed mental models of how encrypted email provides privacy guarantees. Despite the dominant theme of privacy, all participants primarily used Proton Mail to send *unencrypted* email to contacts on other email systems, leading to rather limited privacy gains.

The contributions of our paper include (a) a rich, qualitative data from a set of people who have actively chosen to use a secure email system; (b) analysis of the data that illustrates motivations to use a secure email system, mental models, threat models, and usage patterns; and (c) reflections on how researchers and industry can capitalize on the desire for privacy to realize stronger privacy gains for users.

2 Related Work

Because our work focuses on adoption, we reference several prominent theories from research on technology adoption. The Technology Acceptance Model (TAM) [8] identifies perceived usefulness and perceived ease-of-use as factors influencing behavioral intention to use a technology. The Unified Theory of Acceptance and Use of Technology (UTAUT) [46] extends TAM by considering additional factors such as social influence, voluntariness, and facilitating conditions. Protec-

tion Motivation Theory (PMT) [26, 35] addresses the cognitive processes involved in behavior change when faced with a threat, including assessing threat likelihood and severity, evaluating mitigating action efficacy and cost, and considering self-efficacy.

2.1 Adoption of Secure Technology

Recent work by Zou et al. [55] examined adoption and abandonment of a wide range of security and privacy practices, finding that security practices were more widely adopted than privacy practices. Abu-Salma et al. [1] studied the obstacles to adoption of secure communication tools, discovering that majority of participants did not understand E2EE and primarily adopted them for social reasons rather than security benefits. Story et al. [44] measured the usage of and perceptions about private browsing, VPNs, Tor Browser, ad blockers, and antivirus software. They identified several misconceptions and suggested that interventions surrounding these tools should target well-defined threats and address obstacles to user threat models. Kang et al. [23] interviewed individuals regarding privacy and security risks, identifying that people don't take privacy-protective actions due to lack of concern, actions being costly or difficult, and limited knowledge. Other studies have focused on the adoption of individual tools, such as private browsing [13, 18] and VPNs [10, 29], suggesting similar results.

Prior research has also looked into the adoption of 2FA and password managers, finding that usability issues are an obstacle [6, 7], and that stories encouraged people to be willing to adopt 2FA [12]. Other studies have also found evidence that perceived usability issues may not be as significant as misconceptions surrounding 2FA [5].

Regarding password managers (PMs), prior work has found lack of awareness to be a strong reason for non-adoption [2], and that users of built-in PMs are driven by convenience, whereas users of separately installed password managers prioritize security [31]. Mayer et al. [28] discovered that PM adoption in a university setting is largely driven by perceived ease-of-use.

Two studies have examined adoption of secure email. Gaw et al. [16] found that the perception of encryption behaviour by others influenced a person's decision to adopt encrypted email. Renaud et al. [34] found that misaligned incentives, lack of understanding of the email architecture, and fragmented threat models cause the non-adoption of E2E-encrypted email.

2.2 Privacy Frameworks

One of the motivations we found for people adopting secure email was a desire for privacy. Accordingly, we review the variety of theoretical approaches that researchers have used to explain how people conceptualize and treat privacy.

Westin's taxonomy of privacy classifies individuals based on their varying levels of privacy concerns [21, 48]. However, this classification is far from modern real-world scenarios [51] and does not take into account the wider range of privacy management strategies by users [25, 50]. Malhotra et al.'s information privacy concern scale looks at privacy from the perspective of the collection, control and awareness of information [27]. Prior research has also highlighted privacy calculus, in which individuals weigh the costs and benefits of disclosing their personal information [20, 24]. Another prominent privacy framework is contextual integrity [30], that takes into account the social and cultural norms of specific contexts and argues that privacy is maintained when information flows align with these norms.

Solove proposed a taxonomy of privacy threats which includes four categories: information collection, information processing, information dissemination, and invasions [43]. Solove also worked on conceptualizing privacy [42], which takes into account that individuals are likely to differ in their perceptions of what privacy constitutes, how privacy can be violated, and which privacy benefits are most important to them. In this work, he characterized privacy as six major conceptions: (1) the right to be left alone, (2) limited access to self, (3) secrecy, (4) control over information, (5) personhood, and (6) intimacy.

Our findings on privacy motivations for adoption do not align with any singular privacy framework; we discuss this in Section 5.1.

3 Methodology

Our study is focused on the unique population that has decided to *voluntarily* adopt a secure email service. We designed and conducted semi-structured interviews with 25 users of Proton Mail and Tutanota, two popular secure email systems that claim to have 70 million users and several million users, respectively. We used a semi-structured interview guide to ensure we covered material relevant to each of our research questions, while also having the freedom to explore topics in more depth as needed.

3.1 Screening Survey

In all recruiting venues we asked participants to take a short screening survey to confirm their eligibility (age 18 or older, able to speak English), provide a list of email services they have accounts with, indicate the amount of time they have had a secure email address, describe the frequency with which they use their secure email account, and answer basic demographic questions.

Based on results from the screening survey, we used purposive sampling to ensure that we recruited participants who used secure email services across a variety of characteristics such as the amount of time they have been the service for,

how often they use it and whether they use it as their primary email account.

3.2 Recruitment

After substantial recruiting efforts, we were able to recruit eight participants from Reddit and 17 participants from Prolific. We paid participants from Reddit USD 15 each using Amazon gift cards, and participants from Prolific USD 25 each as a Prolific bonus. We increased the compensation for Prolific participants since they were unwilling to participate in a lengthy interview for only USD 15.

Recruiting was challenging because we wanted to interview people who used secure email systems, and this is a relative minority of the overall population with no easy way to access them. We detail some of these challenges below to aid future researchers with similar problems.

We initially posted the invitation for our study on the official subreddits for Proton Mail[2] and Tutanota[3]. After having mixed success, with most participants being technically savvy, we attempted to diversify our sample. We posted our study on Amazon Mechanical Turk and on several general subreddits that were not related to technology. We did not screen for location as long as the potential participants could communicate in English. We asked a few questions at the beginning of the interview to filter fraudulent attempts at participation by non-users, including asking for their zip code (which would typically not match what they had entered in the screening survey), asking for their Proton Mail email address, sending out a test email, and asking about features of Proton Mail that only a user would know. *None* of the participants from MTurk seemed to be legitimate users. We believe the attempt to participate was largely due to the monetary incentive offered, especially in countries with higher USD value, leading in a disproportionate representation of non-users attempting to participate solely for the reward. We therefore decided to exclude MTurk and general subreddits from our study.

We also placed a Google Ad for our study that appeared in search results for terms related to secure and private email, and experimented with both a USD 25 payment for an interview and a drawing for USD 100 with a 1 in 5 chance of winning. Despite the ad receiving 63.5k impressions and 996 clicks, for a total cost of USD 176, nobody signed up for an interview in this recruitment channel.

Ultimately, we switched our recruiting efforts to Prolific, where we had much better success. To mitigate the issue of having non-users in the study, we excluded countries where the ratio of English speakers was extremely low or the currency difference was especially higher. We did not have to exclude any Prolific participants during screening.

[2] https://www.reddit.com/r/ProtonMail/
[3] https://www.reddit.com/r/Tutanota/

3.3 Demographics

We interviewed 25 users of Proton Mail. Participants were residents of Australia, Canada, Mexico, the Netherlands, Switzerland, Portugal, Poland, Spain, Greece, Japan, the United Kingdom, and the United States. Three of them identified as female and 19 identified as male, two identified as non-binary and one preferred not to answer. Four were between 18–24 years of age, twelve were 25–34, four were 35–44, and five were 45–54. Most users were highly educated: nine had bachelor's degrees, and eleven had graduate or professional degrees. 12 participants had a formal background in technical fields. We provide detailed demographics in Table 1.

3.4 Interviews

We conducted all interviews in English remotely via Zoom, where turning the camera on was optional for the participants. Each interview lasted between 35-45 minutes. We began by asking some ice breaker questions to put them at ease, and we confirmed that the participant currently used Proton Mail or Tutanota. To avoid bias, we made sure to not use the word 'security' or 'privacy' until the participant mentioned it. We then asked questions in four different areas, in order, corresponding to each of our research questions:

- *Adoption*: We asked how they first heard about Proton Mail, how they started using it, why they currently use it, whether they encourage other people to use it, and similar questions.
- *Threat model*: We asked them which entities they feel would access or misuse their email data if they could get it, what the consequences would be of someone reading their email without permission, and how they mitigate any perceived threats.
- *Mental model*: We asked participants how they think Proton Mail works. We then asked them to draw what is involved when one person sends an email to another person, similar to prior work [22, 23, 52]. We encouraged participants to think aloud while drawing to gather additional insights into their reasoning. We asked the participants to send a photo of their drawing to us, or if they had their camera on, we requested them to hold it up to the camera and took a screenshot. We explored both structural properties, which describe how participants view the internals of the working of Proton Mail, as well as functional properties which focus on how these users interact with and use the email system.
- *Usage*: We asked them what they use their Proton Mail account for, how they interact with people who don't have secure email accounts, and what they like and dislike about Proton Mail.



Figure 1: Terminology used to convey relative frequency of themes

3.5 Data Analysis

We recorded the audio from each interview using Zoom. We then transcribed the recordings using an automated transcription service. The first author reviewed all transcripts to ensure consistency with the recordings.

We conducted qualitative coding regularly throughout the interview process. This enabled us to look for saturation and to adjust the interviews as interesting ideas or themes emerged. We used thematic analysis, coding the data corresponding to our research questions. We primarily assigned the codes inductively, but used deductive coding for threat models, where we looked specifically for attackers, harms an attacker can cause, and how the participant explained they would mitigate that harm.

Three researchers coded all the transcripts together and disagreements were resolved through consensus-building as they emerged. We started by coding the data, assigning first-order codes which were closely aligned with the terms used by the interviewees in order to preserve the authenticity of their expressions. We then refined the codes through further iterative rounds of analysis, assigning second-order themes [17]. Similar themes were merged together to identify relationships and patterns in the data.

The primary author conducted a separate analysis of the drawings and the accompanying verbal explanations. In doing so, we grouped similar drawings and mental models together based on a participant's understanding of the inner workings of secure email systems. These categories were then reviewed and discussed among all the authors and any discrepancies were reconciled.

Since our work is qualitative in nature, we avoid using exact numbers. Instead, we use a consistent terminology to convey the relative frequency of major themes, as done by previous studies [11, 19, 53]. Figure 1 presents the terms used to indicate the frequency of occurrence of participants' responses.

3.6 Ethical Considerations

Our study did not create significant potential for harm to participants because we only sought to gather their opinions and experiences. The Institutional Review Board (IRB) at Brigham Young University reviewed and approved our study, and we obtained informed consent from participants. Because participants were from a variety of countries, each potentially with their own privacy laws, we took care to notify all participants of their data privacy rights, using a superset of all rights available in countries whose privacy laws are tracked at the Global Data Privacy & Security Handbook.[4] Specifically, we informed all participants that they had the right to access their own data, correct their data where inaccurate or incomplete, erase their personal data, withdraw consent, etc.

3.7 Limitations

We chose an interview study to gain insights into the attitudes and experiences of a relatively understudied group. As with most qualitative work, our purpose was to surface primary themes that impact adoption, understanding, and use of secure email, rather than to quantify the prevalence of these themes. Our sample is diverse among age, location, and technical expertise, but doesn't capture all possible opinions or experiences.

Despite trying to find users of a variety of secure email systems, with a focus on voluntary adoption rather than mandated corporate use, all of our participants primarily used Proton Mail as a secure email system. Further, we interviewed participants who were fluent in English and resided in countries where the currency exchange rate difference with USD was not dramatically high. Thus our results may not reflect the broader secure email space.

4 Findings

In this section, we present the themes we observed across our interviews for each of the research questions we study: (1) Why do people voluntarily adopt secure email systems? (2) What threat models do people have, meaning their conception of attackers and the harms they can impose, and what steps do they take to mitigate these harms? (3) What mental models do people have of secure email systems and their capabilities? (4) Do people use the secure email services effectively and what obstacles they encounter?

All of our participants were active users of Proton Mail (with a few also using Tutanota), so our findings repeatedly reference their use of this system in particular.

4.1 Adoption Motivations

We found a variety of factors that drive the adoption of Proton Mail for our participants. We describe them here in order of their prevalence and level of emphasis.

Distrust of Big Tech: The decision to adopt Proton Mail was driven heavily by the distrust our participants showed toward technology giants. Majority of the participants expressed

[4] https://resourcehub.bakermckenzie.com/en/resources/data-privacy-security

concerns regarding the continuous monitoring and data collection practices employed by these organizations. Participants mentioned feeling being exploited by big tech companies and feeling uncomfortable with companies knowing everything about them, from their location to their interests to what they are purchasing. They reported these surveillance acts as "creepy" (R5) and these companies as "nasty" (R9).

The participants in the study expressed a significant degree of mistrust in the practices of Google and Facebook in particular, viewing their monitoring activities as intrusive and invasive:

> *"Over the course of the last 20 years working on the internet, I have noticed an increasing amount of activity from business entities like Google, that can only be described as creepy. The fact that Google and Facebook and other big corporations like that are able to put together so much information about us as individuals, and take advantage of that to commercially exploit it, and not even give us a cut of the profits." (R5)*

Participants raised concerns about the integration of Google's products, which they believed gave the company comprehensive access to their personal information and ability to profile and track users. Participants likewise mentioned Facebook and its ability to track and share data outside of their own site.

> *"Whatever it is that you put into your computer or your smartphone, it can be seen and it can be listened to... Facebook used to be fun, and then it destroyed democracy. So later, it stopped being fun at a certain point... And I don't feel very comfortable anymore with these companies." (R9)*

They stated that they abstain from using social media as much as they can, and in some cases, entirely, believing that the cost of disclosing information outweighed the benefits. Yet even this was sometimes considered ineffective, given the tracking that these companies use even on non-users of the site.

Participants overall had a general perception that the big tech companies are not conscientious and ethical. This led to a desire to avoid big tech companies whenever possible and choosing a product that offered them more privacy. As one participant put it,

> *"Over the past few years, I've been trying to wean myself off of Google and other, you know, big tech products, Because they are kind of, I think they're poisoning my mind." (R1)*

Privacy: Privacy is also a significant motivating factor for the adoption of Proton Mail among our participants. We characterize the different models of privacy our participants described according to their conceptualizations, similar to [43]. We found that our participants had different conceptions of privacy which sometimes overlapped. Below, we outline these models and provide examples of how they influence the participants' usage of Proton Mail.

Privacy as a fundamental right: Some participants felt that individuals have an inherent and inalienable right to privacy, and that privacy is not just a preference or a convenience, but is instead a necessity.

> *"I fully believe that privacy should be the default on the internet. It's heinous how we've let that completely fall apart. I'm appreciative of the GDPR and everything that it does... But at least in this country (USA), it's pretty much understood that you're the product if you're using the internet. The internet used to be so cool, and now it's just kind of a garbage fire." (R1)*

Privacy as Anonymity: Some participants believe individuals have the right to use the internet and other digital services without revealing their true identity or personally identifiable information. They adopted Proton Mail because it does not require them to enter their phone number in order to create an account. They can *choose* to provide it for account recovery and two-factor authentication but Proton Mail does not impose this on them. They also use pseudonyms on Proton Mail instead of their real names and like the idea that their communications and activities through that account cannot be traced back to their other email accounts. Participants also reported that they liked the fact that Proton Mail did not log their IP addresses unless they activated this feature.

Privacy as Control: Some participants felt that individuals have the right to control the collection, use, and dissemination of their information. Participants with this model mostly used Proton Mail as a secondary, separate account from their main email address, and used it for a specific task that they wanted to not be associated with their primary online identity. This way, they control the information that is associated with each account, and they are able to ensure that the information they want to keep private is only associated with their secondary email account, which often is an account that uses a pseudonym with no personally identifiable information attached to it.

Privacy as Commodity: Some participants viewed privacy as a commodity that can be bought and sold in the marketplace [41]. Some participants with this conception were particularly uncomfortable with the idea that big tech companies are taking their data and using it to their own benefit without giving any benefit to the individual the data belongs to. Others stated that they were exchanging their privacy for the services they were receiving through these tech giants.

Privacy as Secrecy: Some participants based privacy on the principle of confidentiality. They reported using Proton Mail for its encryption properties that prevent Proton from reading a user's emails.

However, not all our users understood this property. Some of them incorrectly believed that even if the emails are encrypted, it protects them from outside attacks but Proton Mail can still see all their communications. Even with this model, they believed that Proton Mail provided them with a higher level of privacy as opposed to an ordinary service, because their information could be seen only by Proton Mail and was not sold to third parties.

While the overall sentiment our participants shared was that all information deserved to be "safe" and "protected", they repeatedly mentioned that since they were not a high-profile personality and were not doing anything illegal either, they had "nothing to hide". We investigated how the participants defined and characterized sensitive information. The most recurring definition we saw was any personally identifiable information. Our participants particularly resort to Proton Mail when they require anonymity. Other definitions of sensitive information included financial or bank account details, authentication credentials such as PINs and passwords, location, and race.

Affordances: We viewed the different ways in which users interact with secure email through an affordances perspective [15, 39], broadly meaning the possibilities of ways users employed secure email to achieve their goals. We found that sometimes, our participants adopted Proton Mail for one particular reason and used it for that reason only. For example, P10 and P12 use their Proton Mail accounts for only *receiving* emails about their cryptocurrrency trades.

Another participant, R9 stated that he uses a Proton Mail account with a pseudonym and has it associated with a Facebook account. He then uses the Facebook account for selling items on marketplace and contacting potential customers. This way, his original identity is never exposed and is therefore not at risk.

Similarly, P13 uses a Proton Mail account for different micro-tasking websites and uses a pseudonym for it. In his opinion, since the micro-tasking websites do not *need* to know his real name or identity, he likes to use Proton Mail for it and then his data is not associated with his main accounts.

P18 mentioned that he sometimes needs to access his email account from different locations in the world and sometimes shares his email account with someone in a different part of the world. For him, the security measure by Gmail that tracks all IP addresses which access his account is not a desirable feature. He uses Proton Mail because it does not do so if you have your authentication logging off (which it is, by default).

Aversion to Personalized Advertisements: Aversion to personalized advertisements is also emerged as an important reason behind adoption of an encrypted email system. Some participants mentioned that they noticed Gmail scanning their emails for keywords and using that information to display personalized ads related to the content of their emails.

Some participants had experience in careers that exposed them to the kind of information collected about an individual and how that information is shared and used. These participants particularly expressed being uneasy with this practice, leading them to switch to a service like Proton Mail that does not engage in such practices.

Although some participants acknowledged that advertisements are a source of revenue for companies, the majority expressed strong dislike for personalized ads, especially when they originated from unexpected sources. Participants also understood that data was shared to third parties, and that avoiding a given service did not guarantee that the service would have no knowledge of their information. Participants had developed this mental model through personal experiences of seeing targeted ads even when they had not used used a particular service before. A majority of the participants particularly expressed this sentiment with regard to Facebook and Google, stating that anything a person does online is known to these two companies. R1, who is not a Facebook user, mentioned that he uses Proton Mail because he does not want Facebook to know all about his communications even though he does not have a Facebook account.

> *"So I wouldn't want [Facebook] to, you know, somehow manage to sniff my communications. Who doesn't hate advertising? I hate advertising." (R1)*

Trust in Proton: Proton Mail advertises itself as a company that 'protects your privacy'. About half of our participants were unaware of the specific ways their data is protected when using Proton Mail, or ways in which Proton Mail differs from other email providers in terms of its functionality. Despite this lack of understanding, they trusted the company's promise of privacy protection. They either did not know or were not concerned about the encryption of their emails, but rather placed their trust in Proton Mail's commitment to not share or exploit their data. As P14 stated, they trusted the company's reputation for protecting privacy.

> *"I'm assuming that the more privacy focused company wouldn't give away my data." (P14)*

Some participants also expressed trust in Proton Mail due to its location. They had the view that since Proton Mail is founded and based in Switzerland, it provides them a higher level of privacy as they cannot be subjected to surveillance on behalf of US or other intelligence agencies. While Proton Mail claims zero-access encryption, a few participants mistakenly believed that Proton Mail has access to all their email communications. Nevertheless, they felt safe knowing that Proton Mail, being subject to Swiss laws, would not be compelled to release their data to US or EU agencies, even when requested to do so. Similar views were expressed by participants who used Tutanota, which is based in Germany

and similarly protected from having to provide data to the US government.

Conscientiousness: Some participants have adopted Proton Mail because they want to support a conscientious company. In a time where data sharing and revenue generation through advertisements and personal data sales are common, they believe that companies like Proton, which prioritize ethical and conscientious practices, should be supported. Our participants stated that users' support for companies that value ethics are important, even at the cost of certain conveniences or functional advantages. Some participants mentioned purchasing the paid plans for Proton Mail instead of using the free version because it makes them feel good about supporting an ethical company.

> *"I purchased a plus subscription to for Proton Mail, because I like supporting conscientious companies like that. So it's partially the privacy and partially it's feeling good about, you know, being a techno vegan." (P16)*

Exposure to Technology and Negative Experiences: Participants with a previous negative experience with technology cited it to be their reason of adoption of an encrypted email service like Proton Mail. Some participants who had not directly had this experience, but had heard about such incidents also felt motivated to use an encrypted email service, as seen by [32, 33] as well.

Further, some participants indicated that exposure to technology served as a driving factor for them to adopt encrypted email services. Their level of awareness about the potential for privacy violations, whether through education or their career, influenced their level of motivation to protect their privacy, since they better understood the likelihood and extent of harm.

4.2 Threat Models

We prompted the participants to think of any entities that could potentially pose a risk to their email communications. They were instructed to perform a think-aloud exercise to identify and articulate the potential threats. Here we describe the categories of attackers and their respective capabilities, as well as any preventative measures participants use to safeguard themselves against these threats.

4.2.1 Adversaries/Attackers

The adversaries our participants mentioned aligned well with the findings of [1] which found that users perceive three types of adversaries: (1) government agencies, (2) service providers, and (3) anonymous hackers. Our participants additionally differentiated between email service providers and other internet-based companies. Further, our participants often clarified that just because an entity has the ability to cause a harm does not necessarily mean that it actually will ever do so. Only one participant (P10) mentioned the risk of someone physically accessing her devices, but dismissed it saying that it is extremely unlikely.

Government and Intelligence agencies: A prevalent potential threat most of our participants perceive is surveillance by governmental agencies. While they mostly think it is unlikely for their government to spy on them and access their emails, they listed it as a possibility nonetheless. Some of our participants mentioned that The Five Eyes Alliance countries (Canada, Australia, New Zealand, the United Kingdom, and the United States) might be more likely to monitor people's email communications. Although the likelihood of such an event happening was deemed negligible, the potential consequences were described as severe. The participants emphasized that governmental entities wield considerable power and could potentially issue directives to email service providers, requiring the surrender of all relevant data. They also believed that governments typically have back doors to encryption algorithms, a sentiment also expressed by interviewees in [1]. The consequences of such an event occurring were perceived as extremely intense and life-threatening such as ethnic cleansing or political assassination.

Anonymous hackers on the internet: According to a majority of our participants, anonymous hackers on the internet pose a credible threat. Nevertheless, the participants held the view that individual attacks on their data are highly unlikely due to the Big Fish model [47] meaning that they are not a significant or "interesting" target (P14) and therefore no one would target them. Rather, the participants expressed concern about the potential for data breaches by these skilled hackers, and getting unauthorized access of corporate databases, since they had often heard such stories. Such breaches were regarded as a serious threat, given the potential to compromise their financial information, which was considered to be the primary motive for such attacks. P17, shared the following experience:

> *"I have seen that there are forums that sell used accounts, for example, for Spotify or PayPal accounts with money on them. So they mostly do it for financial motives." (P17)*

P16 shared a similar experience where their mother's Grammarly account was accessed by an unauthorized individual who obtained the account credentials through a data breach. One participant provided an additional perspective on the potential consequences of hackers gaining access to email addresses, where they could "spam the user to death" (R6) with unsolicited messages until they become overwhelmed

and unable to effectively manage their inbox. The participant described this outcome as highly likely, citing personal experience as evidence.

Other Email Service Providers: Participants identified email service providers to be a potential threat to the privacy of their email communications. More than half of our participants acknowledge that while these practices constitute an infringement of privacy, they understand the economic incentives that motivate these companies to scan and read their emails. They stated that the email providers do not have any malicious motivations, but just need to earn a profit. They reported being particularly annoyed with companies that "grab their attention" and "reduce them to a number of their quarterly earning calls" (R1). Overall, participants expressed relatively low levels of concerns about email providers looking at their information. They held this view due to their belief that they do not have any sensitive information in their emails. Even when realizing that their emails contain their financial information which they consider to be sensitive, they stated that they trust the email providers to not misuse that information. They mentioned that the biggest threat through email providers is probably just targeted ads. Some participants believed that Proton Mail has similar abilities and can view and scan all their (encrypted) emails for advertising and profiling. They trusted Proton, however, to not do so.

Online companies: Many participants identified companies and services on the internet as a separate and more significant threat than email service providers. Based on their perception, such entities collect data without users' consent. In contrast to email services, which only have access to email contents, internet-based services can collect additional data across various dimensions, such as location, health information, financial information, race, and interests. Participants viewed this type of data collection as more intrusive and in-depth, hence posing a more severe threat to their privacy as well as security.

4.2.2 Mitigation Strategies

Participants were asked to describe the strategies they employed to mitigate the risks they mentioned. As seen earlier, one of the primary strategies for the threats posed by email service providers and online companies in general was to use Proton Mail. This was seen as a way to remove themselves and their data from big tech, to provide privacy, or to align their choices with companies that share similar values.

When asked about how they would send sensitive information, participants did not mention any strategies related to E2EE systems. Instead, they suggested using offline channels, such as sending the information by post or meeting the communication partner in person. One participant (P11) considered SMS to be a more secure alternative to email and recommended its use as a mitigation strategy to safeguard against information leaks. Although he acknowledged that telephone operators and governments could still access his information through SMS, he felt that it was a relatively safer option compared to the entire internet. Some participants recommended using virtual private networks (VPNs) to safeguard their online activities. In addition, some participants suggested avoiding social media altogether to prevent privacy violations on the internet.

When thinking about protecting themselves from the government, participants mentioned that there is essentially no way to escape that. Some participants expressed some confidence in using Proton Mail, given its location in Switzerland, as a mitigation strategy. However, they perceived that governments always have back doors and can gain access to any information they want, even when one is using an E2EE system, and that in the worst-case scenario, the government could resort to force to obtain their information. Some participants indicated they could protect against government surveillance by being a law-abiding citizen.

> *"If the US government or I mean, heck, even the Pakistani government really wanted to see my emails, they probably, worst comes to worst, beat it out of me." (R2)*

4.3 Mental Models

Since our sample was diverse with respect to the technical background our participants had, their mental models varied drastically depending on their technical knowledge. As we reviewed these models, we grouped them into two broad categories: (1) A Safer, More Trustworthy WebMail System, and (2) A Private, Encrypted Email System. We describe these below.

A Safer, More Trustworthy Email System: Participants with this model did not have a complicated model for what Proton Mail, or any encrypted service for that matter, does when a user tries to send an email to another user. For them, Proton Mail worked just like a regular email provider except it was *somehow* safer. Structurally, they imagined that the processing of email is similar for Proton Mail, Gmail, Outlook, or any other provider.

Participants with this model had at best only a vague understanding that Proton Mail used encryption. Some participants with this model did not know that email in Proton Mail could be encrypted, and had not seen or heard the word encryption. Some thought that all email providers use encryption, but somehow Proton Mail was safer. One participant thought that using the paid version of Proton Mail provides even better encryption than the free version, which in turn is better than using a regular email provider.

> *"But with paid Proton Mail, according to them, they're doing something that if someone tries to*

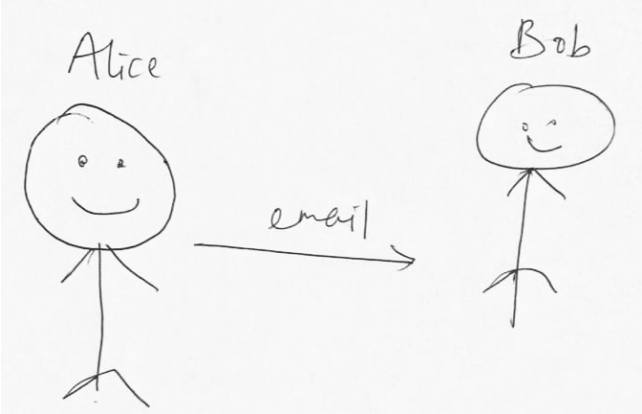


Figure 2: P10's drawing to explain how Alice sends a message to Bob in Proton Mail

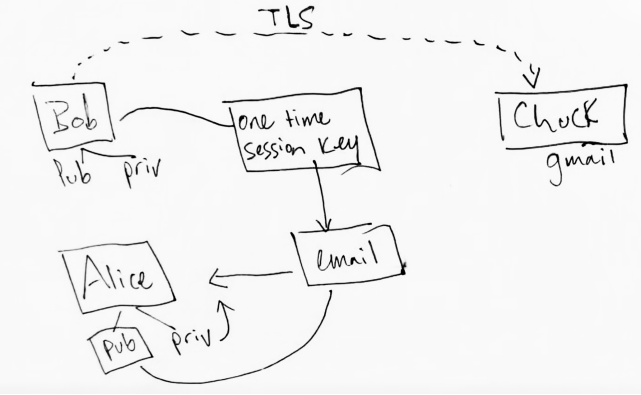


Figure 3: R1's drawing to explain how Bob sends a message to Alice in Proton Mail

> *read the email outside of the system, somehow it's encrypted. I don't know. I don't know how it works. (R6)"*

The common sentiment among participants who held this model is that they do not know Proton Mail works or how it is different than an ordinary email provider, and they probably do not *need* to know the details either. When presented with the diagramming exercise, participants with this model felt at a loss to characterize what goes on in the background when they send an email to their friend. For all they know and care, they send an email and the email is received on the other end *safely*, as Figure 2 shows.

We explored how and why these participants were perceiving Proton Mail to be safer, given that their mental model, both structurally and functionally for Proton Mail and other email providers was essentially identical. We identified that participant perception for Proton Mail originated from the fact that Proton Mail did not collect any personally identifiable information at the time of account creation. While Proton Mail asked them to provide their backup email or phone number for account recovery, this was optional, whereas Gmail and other services they used required those credentials. One participant, P17, mentioned that Proton Mail probably has a better spam filter which makes it safer.

A Private, Encrypted Email System: The other group of participants understood some of the structural properties of Proton Mail and were able to visualize and verbalize the processes involved in sending an email through the system. While some participants made technical errors in describing how encryption works, they generally understood the basic mechanisms.

Participants with this model clearly stated that Proton Mail was different than an ordinary email provider because it is end-to-end encrypted. They also understood that Proton Mail automatically encrypts emails if the sender and receiver are both using Proton Mail, and that emails are encrypted at rest so that Proton can't read them.

Some knew that Proton Mail uses public-key encryption in combination with symmetric encryption. For example as shown in Figure 3, R1 explained this process in detail:

> *"Bob wants to send a message to Alice. If we're talking [about] both Proton Mail users, they both have key pairs. So Bob has a public key and a private key. Alice also has a public key and a private key. And if Bob is the one sending the message, Bob generates a one-time use key. So that's one time, [I] think they call it a session key and uses this key and Alice's public key to encrypt his email. Actually, I should have said, Bob has Alice's public key, [he] uses Alice's public key to encrypt the session key and the one time session key to encrypt the email, which then Alice can decrypt with her private key." (R1)*

Participants with this model clearly distinguished that with ordinary providers, none of these encryption processes are done except that the emails are encrypted in transit through Transport Layer Security protocol (TLS) for security, but that does not protect them from the provider itself because the provider has "all the keys for all the emails". They understood that sending emails from an E2EE email provider to some ordinary provider does not automatically encrypt any emails, whereas encryption automatically happens if both parties use the same E2EE provider. Most participants with this model were aware that Proton Mail provides a password-protected email option that encrypts outgoing emails to someone who is not on Proton Mail. The interviewer hinted at this feature for those who did not mention it themselves. They recalled seeing it but reported almost never using it.

None of the participants with this model mentioned digital

signatures, or address verification. They seemed to trust Proton Mail to distribute the correct keys. They had never seen a warning from Proton Mail about any public key changes for their contacts. They also did not mention the *expiration time* feature for emails sent to other providers, which enables a sender to remove access after a predefined period of time.

4.4 Usage

In this section, we report the ways in which our participants employ end-to-end encrypted email.

What they use it for: About half of our participants mentioned using Proton Mail as their primary personal email account, using it to sen med and receive all personal emails through it. A few participants mentioned using their Proton Mail account exclusively for work and communicating with clients since they perceived their nature of work as sensitive, and that using Proton Mail gave a more creditable look and looked more professional. Some participants mentioned using Proton Mail exclusively for all their communications.

> *"Exclusively for both [work and personal] emails, but in terms of how much time I invest, it's probably around about 75% work and 25% personal." (R7)*

Many of the participants stated using their Proton Mail email addresses as separate, disposable accounts. The main reasons for this are that no personally identifiable information is required to set up an account, thereby simplifying the registration process. Additionally, since these accounts are not linked to their primary online identity, they leverage these 'anonymous' accounts to perform tasks they do not want associated with their main email address. Examples of such tasks include gaming, trading cryptocurrency, completing microtasks on websites such as Prolific, and using Proton Mail as a shared account among multiple in different locations, which they perceived easier due to Proton's no-IP logging policy.

Several participants cited an additional use to exclusively receive newsletters and other superfluous email correspondence, which could otherwise inundate their primary email account. Some participants reported adopting several different email addresses as a means of efficiently managing email content and compartmentalizing them according to distinct purposes, for example using Proton Mail for financial communications, Tutanota for shopping websites, Gmail for everyday usage, and Outlook for school and work-related emails (R8).

Sending to non-Proton Mail users: We asked participants about how they sent emails to contacts who were not using Proton Mail, and their responses indicated that they treated it no differently from sending emails to other Proton Mail users. While some participants mentioned being aware of the password-protected email option offered by Proton Mail, they reported rarely or almost never using it. Even when we hinted at this feature to those who did not mention it, they stated that it was not a feature they ever use. Essentially, our participants are sending and receiving unencrypted emails despite using Proton Mail, since most of their communication partners are not using the platform.

5 Discussion

We didn't seek to validate any general theories of technology adoption. However, TAM seems to broadly apply, since users identify strongly with the usefulness of secure email and current web-based systems have usability roughly similar to popular clients like Gmail. Likewise PMT appears to explain adoption well, since participants have identified specific privacy threats that are highly likely to affect them, Proton Mail offers a reasonable way to mitigate those threats, and they are confident in their ability to use the system. Because these are general theories, they don't adequately capture the broader motivations of our participants, particularly those centered on privacy.

Our study leads to the following takeaways.

5.1 Privacy is a key motivation

In reviewing our findings for each research question, we find that a variety of factors lead to adopting secure email, including distrust of big tech and aversion to the surveillance economy, various notions of privacy, affordances, trust in a company offering these products, and a desire to align decisions with companies that share their values. Privacy permeated many of these motivations.

Privacy also played a role in how participants reacted to perceived threats. Participants who regarded government surveillance as a threat viewed it as highly consequential and potentially life-threatening; however, they did not consider themselves likely targets, and therefore, this was not their primary motivation for adopting encrypted email. Conversely, all participants acknowledged the widespread use of personal data by corporations for targeted advertising, which while a significant invasion of privacy, was not life-threatening. Despite its comparatively lower severity, this threat was more compelling to users, motivating them to adopt ProtonMail.

Furthermore, while security was an added benefit of using encrypted email, it was not primarily security that drove these people to use secure email. Some participants indicated they prioritize privacy over security, preferring Proton Mail because it doesn't ask them for an email or phone number for account verification. Privacy was strongly prevalent among participants who had "A Safer, More Trustworthy Email System" mental model, perhaps because they were unaware of the security threats to their communications and were more exposed to privacy threats.

Although our findings align most closely with Solove's conceptualizations of privacy [42], we did not observe all of the conceptualizations they identified in our research. Moreover, we identified some additional conceptualizations that were not accounted for in Solove's framework. Some participants were highly aware of privacy from the perspective of collection and control of information [27], and some expressed weighing costs and benefits of using a free email system [20, 24]. Thus our participants have diverse understandings of privacy which cannot be easily categorized within a singular privacy framework.

5.2 Privacy benefits are broad

Despite the significant desire for privacy, participants appear to largely be sending unencrypted email to contacts outside of the secure email system they are using. Previous literature has identified inaccurate mental models as a barrier to effective usage of secure technologies [3, 49]. Our results show that even when users possess well-formed mental models with respect to both structural and functional properties, they generally use unencrypted email communication. They understand and are aware that their emails remain unencrypted when communicating with non-users of Proton Mail, which is the case the majority of the time.

The reason for this apparent disconnect is partly rooted in differences in the affordances of secure email systems as viewed by some participants when compared to the expectations of security experts. Many participants found value in pseudonymity (having an email disconnected from their usual account), in avoiding big tech companies, in controlling where their data is stored, or in supporting companies that aligned with their values. Thus privacy benefits are viewed rather broadly, and not tied solely to the ability to send or receive encrypted emails.

5.3 Privacy benefits can be expanded

The relatively low use of encrypted emails among participants does present a significant opportunity for research and industry to *increase* the privacy benefits for secure email users. Future research should explore ways to encourage or nudge users toward password-protecting their emails when sending to users outside the system. There is likely some overlap in methods with research seeking to encourage users of password managers to choose strong passwords instead of storing weak passwords in their password manager [54]. For example, a system could display periodic reminders suggesting emails be encrypted or could display a banner indicating the percent of emails sent in the past week were private.

One clear way to provide greater privacy for existing users is to enable interoperability between secure email systems. Currently, Tutanota does not support PGP, instead uses a proprietary system based on AES and RSA. As a result, it does not automatically recognize and allow importing of public keys attached to an outgoing email from Proton Mail. This prevents users from two large secure email systems from communicating with encrypted emails unless they manually set a password. On the other hand, Proton Mail can exchange secure email with the FlowCrypt Gmail extension, provided the user knows how to attach their public key to an outgoing Proton Mail email, which is not done by default and which is hidden in the user interface behind a menu labeled "..." at the bottom of the compose window. Secure email providers could work together to provide both better support for interoperability and better user experiences for sending encrypted emails. A major challenge is helping users decide whether they should trust another user's key. Trust might be increased by having secure email services automatically retrieve a key for a user from their provider, with that key being signed by the user's email provider.

Ultimately, the best way to provide greater privacy is for secure email systems to have greater numbers of users. Email sent between users of the same system are encrypted by default. One possible avenue is to explore the effect of advertising privacy as the primary feature offered by these systems. Typically marketing literature mixes privacy benefits with promotion of security benefits, while using specialized jargon about encryption. For example, Proton Mail's home page uses the tagline "Secure email that protects your privacy", leading with "secure", and also promotes "independently audited end-to-end encryption and zero-access encryption to secure your communications". Later the home page for Proton Mail explains that encryption "protects against data breaches and ensures no one (not even Proton) can access your inbox". At least some of our users did not notice or understand these benefits. How can industry encourage greater understanding of the benefits of secure email? Would greater awareness and understanding yield more users?

6 Conclusion

Among those we interviewed, privacy concerns are a significant motivator for adopting a secure email system. Web-based systems such as Proton Mail are relatively new options in this space, and participants value the ability to use these accounts to achieve a measure of privacy. These benefits are recognized and appreciated even by those without a deep understanding of encryption, in part because those benefits are significantly broader than traditionally recognized by the security community. Additional research is needed to encourage greater use of encryption, to enable interoperability among providers, and to expand awareness and understanding of the benefits offered by privacy technologies.

Acknowledgments

This research is supported in part by the National Science Foundation under Grant No. CNS-1816929.

References

[1] R. Abu-Salma, M. A. Sasse, J. Bonneau, A. Danilova, A. Naiakshina, and M. Smith. Obstacles to the adoption of secure communication tools. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 137–153, 2017.

[2] Nora Alkaldi and Karen Renaud. Why do people adopt, or reject, smartphone password managers? *1st European Workshop on Usable Security*, pages 1–14, 2016.

[3] Sonia Chiasson, P.C. van Oorschot, and Robert Biddle. A usability study and critique of two password managers. In *15th USENIX Security Symposium (USENIX Security 06)*, Vancouver, B.C. Canada, July 2006. USENIX Association.

[4] Jeremy Clark, Paul C van Oorschot, Scott Ruoti, Kent Seamons, and Daniel Zappala. Sok: Securing email—a stakeholder-based analysis. In *Financial Cryptography and Data Security: 25th International Conference, FC 2021, Virtual Event, March 1–5, 2021, Revised Selected Papers, Part I 25*, pages 360–390. Springer, 2021.

[5] Jessica Colnago, Summer Devlin, Maggie Oates, Chelse Swoopes, Lujo Bauer, Lorrie Cranor, and Nicolas Christin. "it's not actually that horrible" exploring adoption of two-factor authentication at a university. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–11, 2018.

[6] Sanchari Das, Andrew Dingman, and L Jean Camp. Why johnny doesn't use two factor a two-phase usability study of the fido u2f security key. In *Financial Cryptography and Data Security: 22nd International Conference, FC 2018, Nieuwpoort, Curaçao, February 26–March 2, 2018, Revised Selected Papers 22*, pages 160–179. Springer, 2018.

[7] Sanchari Das, Andrew Kim, Ben Jelen, Joshua Streiff, L Jean Camp, and Lesa Huber. Why don't older adults adopt two-factor authentication? In *Proceedings of the 2020 SIGCHI Workshop on Designing Interactions for the Ageing Populations-Addressing Global Challenges*, 2020.

[8] Fred D. Davis. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13(3):319–340, 1989.

[9] Zakir Durumeric, David Adrian, Ariana Mirian, James Kasten, Elie Bursztein, Nicolas Lidzborski, Kurt Thomas, Vijay Eranti, Michael Bailey, and J. Alex Halderman. Neither snow nor rain nor mitm...: An empirical analysis of email delivery security. In *Proceedings of the 2015 Internet Measurement Conference*, IMC '15, page 27–39, New York, NY, USA, 2015. Association for Computing Machinery.

[10] Agnieszka Dutkowska-Zuk, Austin Hounsel, Andre Xiong, Molly Roberts, Brandon Stewart, Marshini Chetty, and Nick Feamster. Understanding how and why university students use virtual private networks. *arXiv preprint arXiv:2002.11834*, 2020.

[11] Pardis Emami-Naeini, Henry Dixon, Yuvraj Agarwal, and Lorrie Faith Cranor. Exploring how privacy and security factor into iot device purchase behavior. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2019.

[12] Chris Fennell and Rick Wash. Do stories help people adopt two-factor authentication? *Studies*, 1(2):3, 2019.

[13] Kevin Gallagher, Sameer Patil, and Nasir Memon. New me: Understanding expert and non-expert perceptions and usage of the tor anonymity network. In *Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017)*, pages 385–398. USENIX Association, 2017.

[14] Simson L. Garfinkel and Robert C. Miller. Johnny 2: A user test of key continuity management with s/mime and outlook express. In *Proceedings of the 2005 Symposium on Usable Privacy and Security*, SOUPS '05, page 13–24, New York, NY, USA, 2005. Association for Computing Machinery.

[15] William W Gaver. Technology affordances. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 79–84, 1991.

[16] Shirley Gaw, Edward W. Felten, and Patricia Fernandez-Kelly. Secrecy, flagging, and paranoia: Adoption criteria in encrypted email. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '06, page 591–600, New York, NY, USA, 2006. Association for Computing Machinery.

[17] Dennis A Gioia, Kevin G Corley, and Aimee L Hamilton. Seeking qualitative rigor in inductive research: Notes on the gioia methodology. *Organizational research methods*, 16(1):15–31, 2013.

[18] Hana Habib, Jessica Colnago, Vidya Gopalakrishnan, Sarah Pearman, Jeremy Thomas, Alessandro Acquisti, Nicolas Christin, and Lorrie Faith Cranor. Away from prying eyes: Analyzing usage and understanding of private browsing. In *Fourteenth symposium on usable*

privacy and security (SOUPS 2018), pages 159–175, 2018.

[19] Hana Habib, Sarah Pearman, Jiamin Wang, Yixin Zou, Alessandro Acquisti, Lorrie Faith Cranor, Norman Sadeh, and Florian Schaub. " it's a scavenger hunt": Usability of websites' opt-out and data deletion choices. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2020.

[20] Il-Horn Hann, Kai-Lung Hui, Tom Lee, and Ivan Png. Online information privacy: Measuring the cost-benefit trade-off. *ICIS 2002 proceedings*, page 1, 2002.

[21] Louis Harris, Alan F Westin, et al. Consumer privacy attitudes: A major shift since 2000 and why, 2003.

[22] David Jonassen and Young Hoan Cho. Externalizing mental models with mindtools. *Understanding models for learning and instruction*, pages 145–159, 2008.

[23] Ruogu Kang, Laura Dabbish, Nathaniel Fruchter, and Sara Kiesler. my data just goes everywhere:" user mental models of the internet and implications for privacy and security. In *Eleventh Symposium on Usable Privacy and Security (SOUPS 2015)*, pages 39–52. Ottawa, 2015.

[24] Robert S Laufer and Maxine Wolfe. Privacy as a concept and a social issue: A multidimensional developmental theory. *Journal of social Issues*, 33(3):22–42, 1977.

[25] Jialiu Lin, Bin Liu, Norman Sadeh, and Jason I Hong. Modeling users' mobile app privacy preferences: Restoring usability in a sea of permission settings. 2014.

[26] James Maddux and Ronald Rogers. Protection motivation and self-efficacy: A revised theory of fear appeals and attitude change. *Journal of Experimental Social Psychology*, 19:469–479, 09 1983.

[27] Naresh K Malhotra, Sung S Kim, and James Agarwal. Internet users' information privacy concerns (iuipc): The construct, the scale, and a causal model. *Information systems research*, 15(4):336–355, 2004.

[28] Peter Mayer, Collins W Munyendo, Michelle L Mazurek, and Adam J Aviv. Why users (don't) use password managers at a large educational institution. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 1849–1866, 2022.

[29] Moses Namara, Daricia Wilkinson, Kelly Caine, and Bart P Knijnenburg. Emotional and practical considerations towards the adoption and abandonment of vpns as a privacy-enhancing technology. *Proceedings on Privacy Enhancing Technologies*, 2020(1):83–102, 2020.

[30] Helen Nissenbaum. Privacy as contextual integrity. *Wash. L. Rev.*, 79:119, 2004.

[31] Sarah Pearman, Shikun Aerin Zhang, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. Why people (don't) use password managers effectively. In *Fifteenth Symposium on Usable Privacy and Security ({SOUPS} 2019)*, pages 319–338, 2019.

[32] Katharina Pfeffer, Alexandra Mai, Edgar Weippl, Emilee Rader, and Katharina Krombholz. Replication: Stories as informal lessons about security. In *Eighteenth Symposium on Usable Privacy and Security (SOUPS 2022)*, pages 1–18, 2022.

[33] Emilee Rader, Rick Wash, and Brandon Brooks. Stories as informal lessons about security. In *Proceedings of the Eighth Symposium on Usable Privacy and Security*, pages 1–17, 2012.

[34] Karen Renaud, Melanie Volkamer, and Arne Renkema-Padmos. Why doesn't jane protect her privacy? In *International Symposium on Privacy Enhancing Technologies Symposium*, pages 244–262. Springer, 2014.

[35] Ronald W. Rogers. A protection motivation theory of fear appeals and attitude change1. *The Journal of Psychology*, 91(1):93–114, 1975. PMID: 28136248.

[36] Scott Ruoti, Jeff Andersen, Luke Dickinson, Scott Heidbrink, Tyler Monson, Mark O'neill, Ken Reese, Brad Spendlove, Elham Vaziripour, Justin Wu, Daniel Zappala, and Kent Seamons. A usability study of four secure email tools using paired participants. *ACM Trans. Priv. Secur.*, 22(2), April 2019.

[37] Scott Ruoti, Jeff Andersen, Travis Hendershot, Daniel Zappala, and Kent Seamons. Private webmail 2.0: Simple and easy-to-use secure email. UIST '16, page 461–472, New York, NY, USA, 2016. Association for Computing Machinery.

[38] Scott Ruoti, Jeff Andersen, Tyler Monson, Daniel Zappala, and Kent Seamons. A comparative usability study of key management in secure email. In *Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018)*, pages 375–394, Baltimore, MD, August 2018. USENIX Association.

[39] Andrea Scarantino. Affordances explained. *Philosophy of Science*, 70(5):949–961, 2003.

[40] Steve Sheng, Levi Broderick, Colleen Alison Koranda, and Jeremy J Hyland. Why johnny still can't encrypt: evaluating the usability of email encryption software. In *Symposium On Usable Privacy and Security*, pages 3–4. ACM, 2006.

[41] H Jeff Smith, Tamara Dinev, and Heng Xu. Information privacy research: an interdisciplinary review. *MIS quarterly*, pages 989–1015, 2011.

[42] Daniel J Solove. Conceptualizing privacy. *California law review*, pages 1087–1155, 2002.

[43] Daniel J Solove. Understanding privacy. 2008.

[44] Peter Story, Daniel Smullen, Yaxing Yao, Alessandro Acquisti, Lorrie Faith Cranor, Norman Sadeh, and Florian Schaub. Awareness, adoption, and misconceptions of web privacy tools. *Proceedings on Privacy Enhancing Technologies*, 2021(3):308–333, 2021.

[45] Christian Stransky, Oliver Wiese, Volker Roth, Yasemin Acar, and Sascha Fahl. 27 years and 81 million opportunities later: Investigating the use of email encryption for an entire university. IEEE Computer Society, May 2022.

[46] Viswanath Venkatesh, Michael G. Morris, Gordon B. Davis, and Fred D. Davis. User acceptance of information technology: Toward a unified view. *MIS Quarterly*, 27(3):425–478, 2003.

[47] Rick Wash. Folk models of home computer security. In *Proceedings of the Sixth Symposium on Usable Privacy and Security*, pages 1–16, 2010.

[48] Alan F Westin et al. The dimensions of privacy: A national opinion research survey of attitudes toward privacy. 1979.

[49] Alma Whitten and J. D. Tygar. Why johnny can't encrypt: A usability evaluation of PGP 5.0. In *8th USENIX Security Symposium (USENIX Security 99)*, 1999.

[50] Pamela Wisniewski, AKM Islam, Heather Richter Lipford, and David C Wilson. Framing and measuring multi-dimensional interpersonal privacy preferences of social networking site users. *Communications of the Association for information systems*, 38(1):10, 2016.

[51] Allison Woodruff, Vasyl Pihur, Sunny Consolvo, Lauren Schmidt, Laura Brandimarte, and Alessandro Acquisti. Would a privacy fundamentalist sell their dna for $1000... if nothing bad happened as a result? the westin categories, behavioral intentions, and consequences. In *Symposium on Usable Privacy and Security (SOUPS)*, volume 5, page 1, 2014.

[52] Justin Wu and Daniel Zappala. When is a tree really a truck? exploring mental models of encryption. In *SOUPS@ USENIX Security Symposium*, pages 395–409, 2018.

[53] Shikun Zhang, Yuanyuan Feng, Yaxing Yao, Lorrie Faith Cranor, and Norman Sadeh. How usable are ios app privacy labels? *UMBC Faculty Collection*, 2022.

[54] Samira Zibaei, Dinah Rinoa Malapaya, Benjamin Mercier, Amirali Salehi-Abari, and Julie Thorpe. Do password managers nudge secure (random) passwords? In *Eighteenth Symposium on Usable Privacy and Security (SOUPS 2022)*, pages 581–597, Boston, MA, August 2022. USENIX Association.

[55] Yixin Zou, Kevin Roundy, Acar Tamersoy, Saurabh Shintre, Johann Roturier, and Florian Schaub. Examining the adoption and abandonment of security, privacy, and identity theft protection practices. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2020.

Appendix

A. Interview Guide

Before we start, I just wanted to say thank you for agreeing to help us with our research project. We really value what you have to say. I also want to be sure you know that there are no right or wrong answers to the questions I'm going to ask. We really just want to hear what you think and feel and hear your opinions. Also, if you're ever confused by a question I'm asking, please let me know, and I'll try to explain or rephrase. I will be recording this interview to transcribe the data. Your video will not be used, and it will be discarded as soon as I get the interview transcribed.

Opening Questions

- Do you have any questions before we start?
- Where do you currently live? How long have you lived there?
- Do you have a CS background? What do you do?
- Verify if they use ProtonMail or Tutanota or not.

Adoption

- How did you first hear about (ProtonMail/Tutanota)?
- Why did you decide to start using (ProtonMail/Tutanota)?
  - Was there a specific event that caused you to use (ProtonMail/Tutanota)?
  - Did you consider using any other encrypted email services?
- Why do you currently use (ProtonMail/Tutanota)?
  - Are there multiple reasons?
  - How would you rank these reasons in order of priority?
  - What kind of information do you regard as "sensitive"? (if applicable)
- What do you particularly like about (ProtonMail/Tutanota)? Dislike about (ProtonMail/Tutanota)?
- Do you use WhatsApp? Signal? Viber? Why or why not?
  - What are the pros of using (ProtonMail/Tutanota) over a more traditional email provider?
  - What are the drawbacks of using (ProtonMail/Tutanota) over a more traditional email provider?
    - Are these sets of pros/cons acceptable?
    - Do any of these contribute to your use of a normal email provider?
- Perception
  - How would you rank yourself on how much you care about security and privacy? On a scale of 1-5?
  - Why is that?
  - How would other people rank you?
- Evangelism
  - Have you ever encouraged your friends to use secure email?
  - Why or why not?
  - What would be the 'talking points' of (ProtonMail/Tutanota) if you were to suggest it to someone?
  - (If yes to above) Do people tell you they are not interested in secure email? What are their reasons? How do you deal with that?
  - Have you ever helped anyone get started with (ProtonMail/Tutanota)? What did they need help with? Tell me about an instance.

Threat model

- Are there entities that you feel would access or misuse your email data if they could get it?
  - Who are they?
  - If you could rank these threats, which are the most likely or most severe?
  - Why do you think they would try to access your information?
- What would be the consequences of someone being able to read your emails?
- What would be the consequences of someone modifying an email you sent?
- What would be the consequences of someone forging an email that was supposedly from you?
- Do you have other accounts that could be compromised if your emails get compromised and read by someone else?

Mental Models

- How do you think (ProtonMail/Tutanota) works?
- Could you draw us a picture of what is involved when a person, Bob, sends an email to another person, Alice, when they are both using (ProtonMail/Tutanota)?
  - How is the email kept secure or private?
- Could you draw another us a picture of what is involved when a person, Bob, sends an email to another person, Alice, but Bob is using (ProtonMail/Tutanota) and Alice is using Gmail?
  - How is the email kept secure or private?
- (ProtonMail/Tutanota) is often advertised as being "secure". What do you think that means?
- (ProtonMail/Tutanota) is also often advertised as offering "privacy"? What do you think that means? How is it different from security?
- Do you feel confident that you know enough about technology to use (ProtonMail/Tutanota) successfully?
- Do you feel a person would need your level of understanding to use (ProtonMail/Tutanota) successfully?

Usage

- What do you use your secure email account for?
  - Do you use it as your primary email account?
  - (if applicable) do you use it for all emails or some emails?
  - If you use a non-secure email account as well, how do you decide which to use and when?
- What features do you wish your secure email service had that are not currently offered?
  - Do you have any difficulties using your secure email service?
  - Can you tell us about one recent instance?
- Do you insist people send you email using a secure email service?
  - If so, how is this received?
- Are there any particular features of (ProtonMail/Tutanota) you really like?
- Can you easily email people who do not use (ProtonMail/Tutanota)?
  - (if not) How much does this affect you on a daily or weekly basis?

  - Would adding this feature be a high priority for you?
- If you need to send sensitive information to someone who is not using (ProtonMail/Tutanota), what do you do?
  - How often does that happen?
- Does it bother you when you have to send emails to non-protonmail users? (because gmail or other service providers still do have access to it)

Ending

- How effective do you think your choice of shifting to secure email has been in protecting your privacy? Especially because most of your friends do not use secure email?
- (if applicable) Don't you think Google can still profile you and see your emails if you send email from ProtonMail to Gmail?
- What other steps do you take to protect your privacy (Other search engines, VPNs, etc?)

B. Participant Demographics

Table 1: Demographics of the interview participants

ID	Age	Country	Gender	Education Level	Tech Background	Using for	Frequency of Usage
R1	35-44	United States	Male	G/PD	Yes	5+ years	Daily
R2	45-54	United States	-	G/PD	Yes	5+ years	Daily
R3	45-54	United States	Male	BA/BS	Yes	5+ years	Weekly
R4	45-54	United States	Male	BA/BS	Yes	5+ years	Weekly
R5	45-54	Australia	Male	G/PD	Yes	5+ years	Daily
R6	45-54	United States	Female	BA/BS	No	5+ years	Daily
R7	25-34	United States	Male	G/PD	No	2-3 years	Weekly
R8	25-34	United States	Male	BA/BS	Yes	5+ years	Monthly
P9	35-44	Canada	Male	G/PD	No	few months	Daily
P10	25-34	Portugal	Female	G/PD	No	1 year	1-2 times a year
P11	18-24	Poland	Male	HS	No	2-3 years	Monthly
P12	35-44	Mexico	Non-Binary	BA/BS	Yes	5+ years	Monthly
P13	25-34	Portugal	Male	BA/BS	No	2-3 years	1-2 times a year
P14	25-34	Netherlands	Male	G/PD	No*	1 year	Weekly
P15	35-44	United Kingdom	Female	G/PD	No	2-3 years	Daily
P16	18-24	Spain	Male	Some college	Yes	1 year	Weekly
P17	25-34	Poland	Male	G/PD	Yes	few months	1-2 times a year
P18	25-34	Mexico	Male	BA/BS	Yes	5+ years	Monthly
P19	25-34	Switzerland	Non-binary	HS	No	5+ years	Daily
P20	25-34	Australia	Male	G/PD	No	5+ years	1-2 times a year
P21	25-34	Greece	Male	G/PD	No	5+ years	Weekly
P22	25-34	Mexico	Male	BA/BS	No	5+ years	Weekly
P23	25-34	Japan	Male	BA/BS	Yes	1 year	Monthly
P24	18-24	Poland	Male	HS	Yes	2-3 years	1-2 times a year
P25	18-24	Poland	Male	Some college	No	1 year	Daily

G/PD = Graduate/Professional Degree
BA/BS = Bachelor's Degree
HS = High School

* P14 mentioned being interested in cybersecurity, but does not have a formal background in it.

"Is Reporting Worth the Sacrifice of Revealing What I've Sent?": Privacy Considerations When Reporting on End-to-End Encrypted Platforms

Leijie Wang
University of Washington

Ruotong Wang
University of Washington

Sterling Williams-Ceci
Cornell University

Sanketh Menda
Cornell Tech

Amy X. Zhang
University of Washington

Abstract

User reporting is an essential component of content moderation on many online platforms—in particular, on end-to-end encrypted (E2EE) messaging platforms where platform operators cannot proactively inspect message contents. However, users' privacy concerns when considering reporting may impede the effectiveness of this strategy in regulating online harassment. In this paper, we conduct interviews with 16 users of E2EE platforms to understand users' mental models of how reporting works and their resultant privacy concerns and considerations surrounding reporting. We find that users expect platforms to store rich longitudinal reporting datasets, recognizing both their promise for better abuse mitigation and the privacy risk that platforms may exploit or fail to protect them. We also find that users have preconceptions about the respective capabilities and risks of moderators at the platform versus community level—for instance, users trust platform moderators more to not abuse their power but think community moderators have more time to attend to reports. These considerations, along with perceived effectiveness of reporting and how to provide sufficient evidence while maintaining privacy, shape how users decide whether, to whom, and how much to report. We conclude with design implications for a more privacy-preserving reporting system on E2EE messaging platforms.

1 Introduction

The emerging threats of online harassment and other offensive behaviors pose significant challenges to online platforms. A 2021 Pew survey found that 41% of Americans reported personally experiencing harassment and bullying online [41]. Despite the deployment of algorithms by online platforms (e.g., Facebook [10], Reddit [17]) to detect abusive messages proactively, user reporting remains a widely used strategy across platforms to tackle online harassment [19]. After users report abusive messages, human moderators review reports and decide whether to sanction the reported user.

Compared to other online platforms, end-to-end encrypted (E2EE) messaging platforms such as WhatsApp must rely more heavily on user reporting to regulate online harassment. As E2EE prevents third parties from accessing conversations without users' consent, platforms cannot deploy algorithms to detect abusive messages proactively unless they use client-side scanning, an approach that violates privacy guarantees [35] and creates new privacy risks for users [1]. Hence, user reporting is considered the most privacy-preserving moderation approach for E2EE platforms [35, 53].

However, while reporting can be used to safeguard a user's privacy in the face of abuse, it also carries privacy risks of its own. On the one hand, reporting helps to protect users' privacy when abusers expose their sensitive information (e.g., intimate photos or sexual orientation) to a broader audience [42, 58]. In such cases, reporting these abusive messages so that they get removed can help prevent further dissemination of users' personal information [61]. On the other hand, user reporting also poses new privacy risks—while reporting does not violate E2EE privacy guarantees as it is user-initiated [35], it may expose private information to platforms and moderators. For instance, if users believe the context around the reported message is shared with moderators, they may hesitate to report if sensitive personal information is exposed within the context [4]. Such situations might arise more frequently in the context of online harassment where the harasser is known to the user, such as an ex-partner or family member [15, 45]. Similarly, journalists might be concerned about exposing metadata such as account and device information in a report, which could disclose the identities of their sources to E2EE platforms, even if this information is legitimately

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.
USENIX Symposium on Usable Privacy and Security (SOUPS) 2023.
August 6–8, 2023, Anaheim, CA, USA

useful for making informed moderation decisions [43, 50].

In this paper, we seek to understand people's privacy concerns when considering reporting on end-to-end encrypted (E2EE) messaging platforms. Inspired by prior research that suggests people's mental models of technologies influence their privacy behaviors [21, 39], we start by investigating people's mental models of user reporting, including their assumptions regarding *data flows*, or what data is shared with E2EE platforms and moderators, and how the data is stored and used by platforms. In particular, we are interested in the following two research questions.

RQ1 What are users' mental models of reporting unwanted messages on E2EE messaging platforms?

RQ2 What privacy concerns and considerations do users have when they make reporting decisions on E2EE messaging platforms?

We conducted 16 semi-structured interviews with users of E2EE platforms. To help users articulate their mental models of reporting, we provided participants with cards labeled with stakeholders (e.g., platform moderators, community moderators), data (e.g., reported message, account information), and moderation actions (e.g., delete messages, ban account). We then invited them to organize these cards on a digital board to illustrate a reporting procedure while speaking aloud. As users' privacy considerations are often grounded in their reporting decisions, we created a series of hypothetical scenarios involving abusive messages that also expose different kinds of personal information (Fig. 2) and asked participants to talk through their reporting decisions in these scenarios.

We find that participants assume platforms already collect account and device information and that many platforms also store longitudinal report history and some context for reported messages. While participants believe that platforms will use data from user reports to build more sophisticated anti-harassment tools, they also worry that platforms may misuse this data in ways that benefit the platform at the cost of users' privacy. We also observe that participants have assumptions about the respective strengths and risks of platform moderators versus community moderators. Platform moderators are assumed to be more distant and professional, and are thus more trusted with private information. Meanwhile, community moderators are assumed to be more familiar and therefore present greater privacy risks, even as they may also have more context and more time to properly address reports.

Finally, we find participants make nuanced decisions about whether, to whom, and how much to report based on trade-offs between privacy risks and protections. When reporting, participants want to share just enough information for moderators to make informed decisions, though they are willing to share more if the report is anonymized or they trust platforms. But sometimes reporting is perceived as too much of a sacrifice in terms of privacy for too little gain. Based on these findings, we argue that a more privacy-preserving reporting system on E2EE platforms should provide more granularity for users to tailor their reports, enable more flexible interaction between stakeholders to ensure informed procedures, and have more transparency to cultivate users' mental models of reporting.

2 Related Work

2.1 Combating online harassment and hate with user reporting

Considerable research has established the pervasiveness of online harassment and hate for users of social media, particularly those from marginalized communities [7, 41, 59]. A Pew survey in 2017 found that 41% of Americans reported personally experiencing varying degrees of harassment and bullying online [41]. These abusive behaviors range from trolling that intentionally provokes audiences with inflammatory remarks [18] to "SWATing," in which attackers falsely report emergencies to send police to the target's address [34].

User reporting is an essential defense against online harassment and hate [59]. In this work, we distinguish between reporting to platform moderators versus to community moderators. Here, *communities* refer to groups of multiple chat rooms (e.g., a Matrix community [48] or a WhatsApp community [63], or in the non-E2EE setting, a Slack workspace [57] or a Discord server [24]). At the platform level, nearly all platforms maintain reporting systems that enable users to send unwanted messages to platform moderators, who are employed to review user reports and make platform-wide moderation decisions according to platform policies regarding impermissible content [6, 54]. At the community level, each community can establish ad-hoc ways to receive user reports. For example, on Discord, users may report to community moderators via direct messages, dedicated channels, or emails [25]. Community moderators are often community members elected or appointed to make community-specific moderation actions in accordance with community guidelines about (un)favorable behaviors [12, 40]. Crawford and Gillespie argue that user reporting represents interactions between users, platforms, algorithms, and broader political forces [19]. In our project, we delve deeper into these interactions by exploring users' perceptions of how data flows from users to communities, platforms, and moderators in reporting systems.

Different platforms have implemented different forms of reporting procedures and data records. For instance, a recent survey found that most online platforms utilize both account information (e.g., email address, account username, and the frequency of account actions) and device information (e.g., IP address) for content moderation [53]. In the crowdsourced moderation system of League of Legends, moderators have access to the entire chat log during the match but players' handles and social contacts are removed to protect privacy [12]. Similarly, on Reddit, the identity of the reporter is kept anonymous to community moderators but known to platform ad-

mins [31]. In contrast, moderators on E2EE platforms have more restricted access to chats. For example, the content of the reported message is not disclosed to moderators on Signal and Matrix [49]. On WhatsApp, reporting an account forwards the last five messages from one's conversation with them to moderators [64], while this number is 30 for Facebook Messenger [14]. However, reporting systems have overall been criticized for being opaque [38]. It is unclear how much of the public information about reporting systems is known to users or what they imagine happens when they submit a report; these questions form the starting point for our study.

While it may be more privacy-preserving to limit the amount of information shared in a user report, additional context can be important for moderators to determine a course of action. Indeed, the user reporting system itself can be co-opted to further abuse [47]. For instance, some platforms hide content that has received many reports until a moderator can review it, and moderators may also be convinced to take content down if enough users have reported it – this can motivate groups to silence others by mass reporting content they dislike [65]. Bad-faith reporters can also try to distort information in their reports. Finally, reports can be used to abuse moderators or waste their time. For instance, community moderators on Reddit receive anonymous reports, opening them to harassment with little risk of sanction [31]. We grapple with the trade-offs in preserving user privacy or allowing users to customize what they share in a report in our Discussion 5.3.

2.2 User reporting on E2EE platforms

E2EE messaging platforms like WhatsApp, Signal, or iMessage are popular among people for private communication. Much like non-E2EE social platforms, online harassment is also a problem on E2EE platforms. A 2022 survey by the ADL found that 12% of adults and 15% of teens have experienced harassment on WhatsApp [16], one of the most common E2EE messaging apps. E2EE platforms rely more on user reporting to regulate online harassment. Despite the increasing use of algorithms to proactively detect abusive messages across non-E2EE platforms [10, 17, 32], the lack of access to messages in E2EE conversations without users' consent makes algorithmic detection impossible [9]. Thus, user reporting in E2EE settings is considered a crucial moderation approach to keep platforms alerted to abuse while still preserving privacy and security [35, 53]. To enable user reporting on E2EE platforms, cryptographic protocols such as message franking [26, 27] allow platform moderators to verify that the sender sent the reported message while also providing deniability to entities other than moderators [60].

2.2.1 Privacy risks of user reporting on E2EE platforms

However, it turns out that 47% of people who have experienced harassment do not bother to report it [16]. While prior research indicates perceived ineffectiveness as one of the primary reasons why people fail to report abuse [38], in this work, we highlight privacy as another important but often neglected factor, especially for users on E2EE platforms who are more invested in maintaining privacy [35, 53].

First, people may be reluctant to report due to private information revealed in the course of a conversation where the harassment occurred. For instance, the context around the reported message might include personally identifiable information or political views that people are unwilling to disclose to third parties [4, 5]. These situations arise more frequently when a harasser is known to the recipient. A Pew survey found that nearly half of Americans (46%) who have experienced online harassment say they know the harasser, including acquaintances (26%), family members (11%), and ex-romantic partners (7%) [15]. Indeed, E2EE messaging platforms are often used for conversations among people who know each other and for sharing sensitive information.

In addition to privacy concerns around platform access to private information that can get leaked, sold, or shared, users may also need to consider the privacy risks of moderators as attackers [59]. With privileged access to private information in reports, moderators could carry out attacks like "doxxing" where targets' personal information (e.g., sexual identity and intimate photos) is exposed to a broader audience [42, 56, 58], or surveillance where targets' devices or accounts are compromised for monitoring purposes [29]. Luca et al. observed that users on E2EE platforms are more worried about the leakage of their sensitive information to people they know than to unknown entities [21]. As users have more interactions with community moderators than platform moderators, they may have more privacy concerns about sharing information with community moderators via reporting.

2.2.2 Mental models of user reporting

Prior work has suggested how users' mental models of technologies may influence their privacy behaviors and concerns [21, 39]. In this work, we also observe users struggling to understand what data is shared and how it is stored and used during the reporting process. This may be due to the degree to which online platforms maintain opaque data policies about their reporting systems [38]. In addition, users are shielded from the decision-making process, with little insight into how moderators use information shared to reach a decision, or even whether they actually make a decision [19].

Further, a limited understanding of E2EE potentially complicates users' mental models of how reporting works on E2EE platforms. Previous studies have shown that users lack confidence and accuracy in their mental models of E2EE platforms [2, 55]. For example, Abu-Salma et al. found that a considerable number of users believe that landline phone calls are not less secure than E2EE communications [3] and that their E2EE communications are vulnerable to eavesdropping by determined attackers [30, 52].

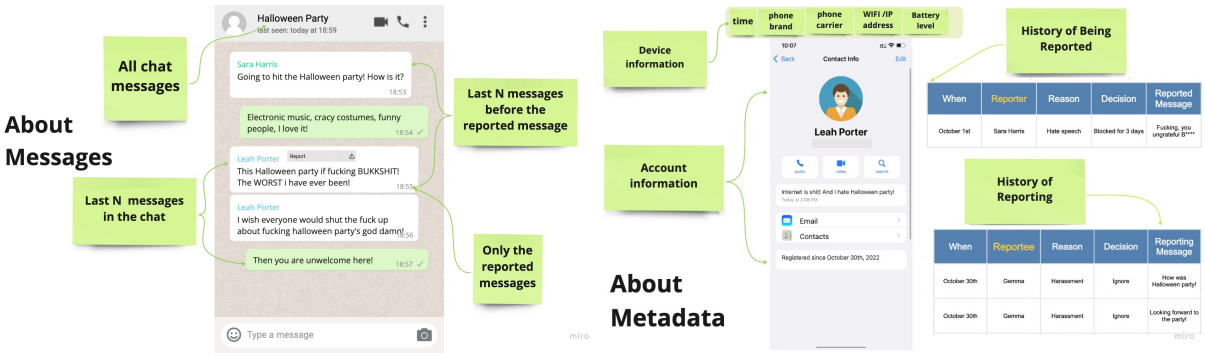


Figure 1: **Mock messaging interfaces annotated with cards to illustrate concepts and necessary background information.** The left board introduced to participants different parts of messages that might be shared with E2EE platforms, whereas the right board lists important metadata that might help moderation, including account information (e.g., registration time, phone number, email address), device information, and history of reporting and being reported.

3 Methods

3.1 Study design and procedures

To understand users' mental models of reporting and their privacy concerns about reporting on E2EE platforms, we conducted semi-structured interviews with active users of E2EE platforms. The final interview protocol was designed iteratively through four pilot interviews to ensure effective elicitation of participants' mental models and contextual concerns. This study was reviewed by our IRB and deemed exempt.

We started the interviews by briefing participants with an overview of the interview session and warning about the possibility of seeing harassment scenarios as part of the interview. We emphasized to participants that they could opt out of questions or stop the interview whenever they wanted, and we gained their explicit consent before we proceeded. We also encouraged participants to think aloud throughout the process. The interview consisted of two sections as described below. The detailed interview protocol can be found in Appendix A.

Section I: Mental models of reporting. In the first section, participants were invited to explain their mental models about how reporting works on E2EE platforms. Inspired by prior work that also investigated mental models [11, 36, 37], we used an interactive card sorting method to better elicit users' mental models. We created a mock messaging interface based on WhatsApp (Fig. 1) annotated with cards to help users get familiar with different components relevant to a reporting system. We also created cards labeled with different concepts (the full set of cards are shown in Appendix A, Fig. 4) on an interactive digital board where users could move around cards to explain their mental models. The digital board prompted participants to reflect on their mental models regarding the following questions: which *stakeholders* have access to reports (e.g., platform, platform moderators, community moderators, hackers, etc.), which *data* is shared (e.g., the reported message, last N messages before the reported message, history of reporting, device information, etc.), and what *moderation action* can be taken (e.g., delete accounts, ban accounts, and delete messages). As participants' perceptions of potential stakeholders are also part of their mental models, we first asked users to explain their understanding of E2EE platforms and to name stakeholders before the card-sorting task.

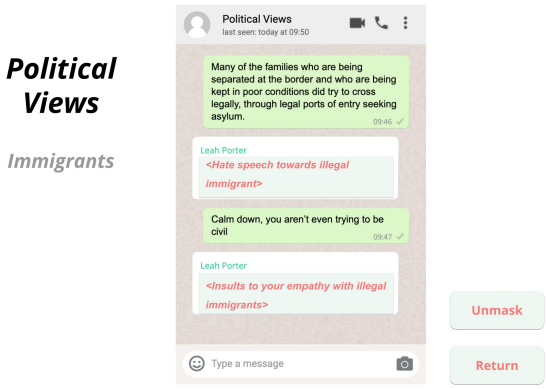


Figure 2: **Example of a hypothetical harassment scenario.** During the interview, participants were encouraged to select information items that they relate to and are comfortable discussing. Here we present a harassment scenario regarding political views. Note abusive languages are masked by default to protect participants from unnecessary harm.

Section II: Privacy considerations about reporting. As not all of our participants have direct experience with online harassment, we first asked them to consider a series of *hypothetical* harassment scenarios (Fig. 2) and decide whether and to whom they are going to report, as well as which information they would like to share with moderators. We designed a variety of harassment scenarios that each had a different type of personal information exposed in the context (the full set

ID	Frequency of Unwanted messages	Reporting Experience	E2EE Platforms	Computer Literacy	Gender	Race
P1	Every few months	N	Telegram	Very high	Man	Asian
P2	Every few weeks	Y	WhatsApp	Medium	Woman	Asian
P3	About weekly	Y	WhatsApp, Messenger	Low	Man	White
P4	Every few months	Y	Signal, WhatsApp	High	Man	White
P5	Every few months	Y	WhatsApp, iMessage	Medium	Man	Asian
P6	Almost never	Y	Signal, WhatsApp	Very high	-	White
P7	Every few months	N	Messenger, iMessage	Medium	Man	-
P8	Almost never	Y	Signal	Very high	Man	White
P9	Every few months	Y	Signal, iMessage	Very high	Woman	White
P10	Almost never	N	Signal	Medium	-	-
P11	Almost never	N	Signal	Very high	Man	White
P12	About weekly	Y	WhatsApp, iMessage	Low	Woman	Asian
P13	Every few weeks	Y	WhatsApp, Signal	High	Woman	Asian
P14	Almost never	N	Signal	Very high	Woman	-
P15	Every few months	Y	iMessage	High	Man	Asian
P16	Every few weeks	Y	Matrix	Very high	-	-

Table 1: **Participant Summary (N=16)**. A single dash means that the participant preferred not to reveal their demographic information.

of scenarios are shown in Appendix A, Fig. 5) so that participants with different experiences can pick scenarios they could better relate to and that they were comfortable viewing [20]. Each scenario is structured around the user first revealing some personal information in a message, followed by an exchange where they receive abusive messages related to that information. We selected different types of personal information from perceived associated risks found by Milne et al. [51]. We then drafted harassment scenarios for each type by first drawing from datasets of hate speech [62] and conversation threads on Twitter. We further iteratively improved our scenarios via feedback from members of our research lab and pilot interviewees to make them sound more realistic. In order to protect participants from unnecessary exposure to traumatizing content, we masked the abusive texts in each scenario with a high-level description and only unmasked them if participants requested.

3.2 Recruitment and participants

We recruited 16 participants for semi-structured interviews who are active users of E2EE group messaging platforms and preferably have reported unwanted messages on these platforms (See Table 1 for detailed demographics). We recruited participants by sharing a recruiting message and a screener survey on Twitter, Mastodon, university-affiliated Slack communities, and privacy-related subreddits including r/europrivacy, r/signal, r/whatsapp, and r/PrivacyGuides.

One challenge of recruitment was that some of our target population are very privacy-aware. To ensure that privacy concerns do not prevent potential participants from signing up for our study, we highlighted our steps to preserve participant privacy in recruiting messages and surveys. For instance, providing demographic information was optional. We only collected participants' email address and name in order to provide compensation, as required by our institution. Participants were also allowed to choose their preferred medium for the interview, including a video call with their camera turned off or via messaging. To observe how participants move cards during the interview, we asked participants to share their screens during the video calls or move cards on Google slides during synchronous interviews over chat.

From responses to our screener survey, we selected 16 participants based on their self-described privacy concerns about personal information. We purposefully did not restrict our recruitment to only individuals who have experienced privacy-related online harassment. Instead, we prioritized participants with a diversity of privacy concerns, reporting experiences, and degrees of computer literacy to capture the mental models of a diverse set of people. As a result, we had only a few people who had directly experienced reporting harassment: one participant spoke of a friend who encountered non-consensual imagery, and one participant spoke about receiving political hate speech. We discuss the limitations of our participant sample in Section 6.

We conducted 15 interviews via video calls, where most participants turned their cameras off but had screen sharing on, and 1 via messaging. The interviews lasted 68 minutes on average, and participants were paid $20. We stopped recruitment when we started hearing repetitive themes and observed no significant new themes.

3.3 Data analysis

We analyzed the interview data qualitatively, following the reflexive thematic analysis approach [13] to understand participants' mental models of reporting and privacy considerations about reporting. Reflexive thematic analysis has been widely used in HCI research to understand users' experience, views, as well as factors that influence and shape particular phenomena or processes [13]. During data collection, the first author took detailed debrief notes after each interview documenting emerging themes. The authors then collectively reviewed the debrief notes and discussed themes in weekly group meetings. Recordings were automatically transcribed into text. The first author then open-coded the data on a line-by-line basis, and the remaining authors reviewed the transcripts and added codes. Over 350 codes were generated from the open-coding process. The authors clustered the open codes into high-level themes in a codebook and iteratively improved the codebook through discussion. Some examples of codes are *data access of platforms*, *trust in community vs. platform moderators*, and *whether I should report*. Finally, the authors applied the codes to the data to complete the thematic analysis.

4 Findings

We first discuss users' mental models of reporting on E2EE platforms regarding data from user reports. We find that participants expect a limited view of the reported conversation, account information, and device information are shared with platforms (§ 4.1.1). Moreover, platform moderators (§ 4.1.2) and community moderators (§ 4.1.3) are expected to only have access to data that are important for reviewing reports. We also find that participants have mixed expectations about how securely the data from user reports might be stored (§ 4.2) and used (§ 4.3) by E2EE platforms afterward.

Following this, we describe how users make careful reporting decisions to both protect themselves against the privacy risks of online harassment and mitigate the privacy risks of reporting. In particular, we discover that participants believe that reporting may fail to protect their privacy against abusers (§ 4.4.1). Participants also perceive that platform and community moderators play different roles in protecting their privacy, despite having more trust in platform moderators (§ 4.4.2). Finally, we find that participants are less willing to share personally identifying information but are willing to share more information if they believe reports are anonymized (§ 4.4.3).

4.1 Data Access of Stakeholders

4.1.1 Platform

Messages. Most participants believe that most E2EE messaging platforms have access to the reported message and the contextual messages around it, which they consider to be important for informed moderation decisions. As P3 described, *"the context helps moderators understand what we've been discussing and where the abuse was being perpetrated or where it was taken place."* Compared to social media platforms in general, E2EE platforms are believed to have more limited access to the context of the reported message. As P11 expressed, *"E2EE platforms get access to that message and the context of that message as well as whatever else they would have to do on my account normally...the contextual might be less on an encrypted message platform."*

However, some participants also expect E2EE platforms with the most stringent privacy principles to have no access to the context around the reported message or even the content of the reported message. These participants are also more privacy-conscious and opted to use these platforms after deliberation. P14 told us that *"Signal can never see messages unless you send them a screenshot or something. So my preference would be that they have as little access as possible."* These participants attribute their expectation to the E2EE platform's reputation for privacy protection: *"if somebody finds out that this platform actually shares the last N messages or this entire chat, this press would be too much of a negative thing for the platform [P1]."*

Account and conversation information. Nearly all participants think that, when they report a message, the platform has access to their account information (i.e., their phone number, email address, registration time, and history of reporting and being reported) and relevant conversation information (i.e., who you chatted with, when you chatted, how frequently you chatted with them). But how much account information each E2EE platform collects also varies. For example, P6 believes that platforms with centralized servers can collect more conversation information when relaying messages between ends, while platforms without centralized servers, such as Signal, collect only a little. As P4 said, *"typically they only know when you last logged on and they typically don't even know who sent a message. Of course, they have to know who to deliver the message to. But as I understand, [Signal] knows almost nothing."*

Device information. Most participants expect that most E2EE platforms collect as much device information as possible. P1 explained this more clearly: *"[the platform] might have access to something that don't require extra permissions [such as] battery, model of the phone, advertising, ID. I would expect they would just collect it."*

Almost all participants are not concerned about sharing their account, conversation, and device information (we refer to these as *metadata* in the following analysis) with the platform when reporting unwanted messages for the next two reasons. First, users have low privacy expectations about these kinds of information since most E2EE platforms do not explicitly guarantee the invisibility of metadata to third parties. In fact, most participants believe that platforms have already collected it before any user reporting. P13 expressed such an idea, *"I'm not worried about sharing [metadata] when I*

report, because I've anyway shared all of this information with them by even using their application. So that information is already with them."

Second, while concerned about the sensitivity of metadata, some participants acknowledge its importance for reviewing user reports. For example, device information can be used to identify whether the sender uses a fake phone, or is a bot. As P6 expressed, *"device information can be very useful and very identifying. I would like for them not to have it at all probably, but it can help to know if a person is using a fake phone or is using an Android VM and Virtual Box, if it's a bot spamming misinformation messages."* Similarly, they believe that account information, especially the history of reporting and being reported, is important for reviewing their reports. P5 said: *"So they need account information to sort of build a profile of that person to understand whether they constantly put in fake requests or they constantly report people and that sort of a thing."*

4.1.2 Platform moderators

Participants have a mixed understanding of the relationship between platform moderators and the platform. Some participants consider these two stakeholders as one entity, thereby believing that platform moderators have the same access as the platform to the information. For example, P2 told us that *"I kind of bunch platform and platform moderators into one entity. I kind of already assume that they have that information because they work for the platform."*

In contrast, other participants think that platform moderators have a lower level of access to data than the platform. We observe uncertainties among these participants about which subset of data is shared with platform moderators exactly. P5 described his uncertainty as follows, *"To me, the platform knows everything, but I don't have a clear idea about where platform moderators lie in that spectrum of the amount of information that they can access."* In the following, we discuss information that participants believe platform moderators may have less access to.

Account information. A considerable number of participants think that platform moderators only have access to non-identifying account information, such as the registration time and history of reporting and being reported, but no personal identifiers like emails, phone numbers, or platform handles. For personal identifiers, participants believe that moderators will *"see some kind of encrypted or hashed kind of version of that number, so they can tell different accounts apart, but they can't kind of reverse engineer who it is [P1]."*

The criteria that participants use to make this distinction is whether they think this piece of account information is important for moderators to make moderation decisions. As P5 has summarized clearly, *"The way I'm thinking about [which information is shared] is what information do I need to make a decision about a specific reported message or a group chat or a person."* He then gave an example: the history of reporting and being reported can be used to identify false reporters and frequent abusers, but *"other account information isn't as revealing as these two pieces of information."* Even for the history of reporting and being reported, several participants expect that a platform that really cares about user privacy would only provide platform moderators with some derivative of this history, such as *"the output of a classification algorithm acting on this information [P14]."*

Device information. Similarly, some participants also expect that only anonymized device information is shared with platform moderators, as evidenced by the words of P11, *"I wouldn't think device information, at least nothing uniquely identifiable, but maybe iPhone or Android or something like that."* Further, several participants also believe that platform moderators are only provided with some derivative of device information because they are not able to interpret device information on their own. P6 believed that *"The platform would be able to [use device information to] track people across different applications, but the moderators themselves would not be able to use it to perform the tracking."*

4.1.3 Community moderators

With a hierarchical structure of reporting systems on messaging platforms in mind, most participants believe that community moderators have access to less data than platform moderators when reviewing user reports. For example, some think that community moderators have access to a limited set of account information, such as the registration time and the location country. They may also only know the history of reports in the scope of their community. P11 implicitly made a distinction between the information community moderators should have versus platform moderators during the interview: *"I'd be fine with a community moderator, for example, knowing how old my account is or maybe what country I'm logged in from, but not something like the email recovery address I use or my IP address. Seems like more of a platform moderator type of information."*

This distinction is due to participants' perception that community moderators, as active participants in the community, are more familiar with parties involved in the reported conversation than platform moderators. As a result, the platform provides platform moderators with more information to make up for their disadvantages, or equivalently, refrains from giving community moderators unnecessary access to more information. P5 explicitly talked about his perception of community moderators: *"The way I think of community moderators is like a person within the community that's also sending messages and constantly an active participant in that community...They have enough context to just look at the messages and make a decision, rather than platform moderators who need more technical information to kind of detect this kind of thing."*

4.2 Data storage

Participants expect that the data from user reports will be stored on the platform server for some time. How long the data will be stored depends on local legal requirements, and moderation requirements as moderators need data from previous user reports for future reference. P11 explained his expectations in detail: *"because sometimes the abuse may not be enough to delete the account or ban it so I should be able to act on that information at a later date. I would also imagine for liability reasons they might be required to keep stuff for a certain amount of time."*

Participants have mixed expectations about the protection of data from user reports by E2EE platforms. Despite acknowledging their limited understanding of technical details, some participants trust E2EE platforms to securely store the data at rest. P13 told us that *"the trust I placed not fully based on understanding technical details, but in trusting them to do the security well."* Other participants, who are more knowledgeable about secure practices, expressed concerns about the platforms' ability to protect the data from potential hacking. For example, P9 was worried that, while the data from reports may be encrypted in transit, it has to be decrypted for moderators' review and might not be encrypted at rest. Platforms may also fail to enforce strict access controls. One participant also expressed concerns that *"even if they had the best intentions and wanted to provide encryption, they might do it for text, but not media, whether it's video or images [P13]."*

4.3 Data usage

In addition to the privacy risks resulting from insecure data storage, participants also believe platforms will further use data from user reports after the reporting procedure for the following purposes. While users do not feel uncomfortable about platforms' developing anti-harassment tools or mining usage patterns based on data from user reports, they are more concerned about being profiled for advertising purposes.

Anti-harassment tooling. Some participants believe that data from user reports will be used to improve anti-harassment tools. Platforms can learn patterns of spurious links in reported messages and block reported accounts that send similar links more proactively. P13 expressed her hopes for this purpose: *"But it's not a worry, it's actually a hope that they would use it towards understanding what issues users face better and trying to make it a safer platform just on the whole."*

Data mining. Participants also envision that the platform may use the data from user reports to analyze usage patterns and inform its development priorities. For example, the platform can tell which operating systems users are using from device information and prioritize security measures on popular operating systems. The content of messages can reveal cultural interaction patterns as suggested by P14, *"there's a lot of tension between religious groups in India—you can link to the language that they're using, try to see if there's an imbalance that needs to be considered in how you develop your platform strategy."* Since the data are analyzed in aggregate, participants do not feel uncomfortable about data mining.

Legal investigation. More than half of the participants believe that data from user reports might be relevant to the investigation of illegal acts (such as terrorism, intimate partner violence, and child abuse) and therefore would be requested by law enforcement agencies. For example, the device and account information of the reported person may help law enforcement agencies identify offenders, and the contextual messages in the reported conversation may also be direct evidence of illegal acts. P14 argued further that the history of user reports could be used to undermine testimony by describing the following example: *"let's say a woman goes to the police and says this guy is stalking me...[by] sending me messages on this platform and waiting outside my house. [But the police] find that you haven't reported any of the messages, suggesting you don't really take it seriously."*

User profiling for targeted ads. Participants have opposing views about whether platforms generate user profiles for ad targeting based on data from user reports. Some participants argue that it is technically and economically infeasible for platforms for two reasons. First, they note that personal information in reported conversations is less organized than the phone number or email address, and therefore platforms need to invest huge computational power, disproportional to the benefits they receive from advertisements, to extract personal information. Participants also believe that it is not a comprehensive way to profile users since *"somebody has to either have reported or been reported, and I don't know what percentage of the users are in that group [P10]."*

Other participants believe that it is highly probable for platforms to profile users using data from reports. Platforms can build a granular social graph from conversation information, associate users' metadata with contact information, or even infer their preferences or identities. As P14 suggested, *"if I'm reporting a bunch of homophobia, then the platform could infer that I'm gay with a high probability from that data."*

Participants also respond with mixed feelings if they know E2EE platforms profit off of generating user profiles from their reports. People who have high privacy expectations about E2EE platforms expressed great disappointment, as P8 clearly stated, *"The way I phrase it to people is your information is valuable, it's financially worth something and they're encouraging you to give them this valuable thing for free."* In contrast, participants who use E2EE platforms because of peer influence are more indifferent, as P5 explained, *"let's say even if WhatsApp is using that information, I'm already being profiled on so many other platforms that adding in this little thing wouldn't affect my day-to-day life as much."*

4.4 Privacy considerations about user reporting on E2EE platforms

4.4.1 Whether I should report?

We observed that participants compare the privacy risks of reporting with the protection the report can provide to decide whether they should report, as P4 summarized, *"the question is, is reporting this message or this person worth the sacrifice of revealing whatever I've sent."* For instance, participants tend to report abusive messages if they are in a group chat but tend not to if in a direct message (DM). P2 explained her reasoning in detail: *"For a DM, I don't have to risk myself being identified [by moderators] because of the photo if I don't want to see the photo anymore...I can just delete it on my end and I won't have to see it. But if it's a group chat, then if I delete it on my end, other people can still see it and I don't want other people to still see it. That's why I have to take the risk to report them and have the photo deleted."*

Ineffectiveness. However, reporting is not always effective in protecting users' privacy, especially when abusers have already had access to their personal information. For example, when personal photos are used in abusive messages, taking down the photos from the chat cannot prevent further dissemination since abusers can take screenshots or download the photos, and banning the abuser from chat may even motivate them to share photos in more chats or on other platforms. Similarly, when abusers weaponize people's address or phone number to make threats, participants chose not to report to the platform because *"[the most] the platform can do is to probably remove the particular user from the account, but at end of the day from the chats, [the abuser] could already note down my home address [and] compromise my security [P3]."*

Inefficiency. In addition to effectiveness, the efficiency of moderation actions also influences people's decisions to report. When abusers send personal photos of victims in a group chat, P11 explained his consideration to us: *"it's really a question of time ... maybe there's fewer extra people know if I got the message deleted fast enough, but I would have to assume that the community moderators would be very fast to do that."*

4.4.2 To whom should I report?

Whether moderators can make an effective and efficient decision to protect users' privacy greatly influences participants' decisions about whom they should report to. Participants think that the following three dimensions contribute to the variance between the effectiveness and efficiency of platform moderators and community moderators.

Moderation areas. More than half of the participants believe different groups of moderators are responsible for different kinds of user reports. Community moderators are more familiar with *community norms* and therefore can better detect disrespectful behaviors, while platform moderators are more knowledgeable about *platform policies* and can better identify reports of illegal behaviors. P1 explicitly talked about this contrast: *"My mental model of reporting to community moderators is more of a personal issue like [complaints that] 'I don't like what's being posted'; my mental model of reporting to platform moderators is more of an illegal stuff like posting child pornography, which violates platform rules."*

Therefore, participants tend to report to moderators who they believe can quickly identify the abusiveness of their report and take swift action to protect their privacy. P7 explained his reasoning to us: *"[Abusers] have the potential to do doxxing type activities. So I would want some quick action and it's clearly against stated platform policies. So I would want the platform to take quick action."*

Moderation actions. Most participants think that platform moderators can delete messages or ban accounts from chat and delete accounts from the platform, while community moderators can do everything but delete accounts from the platform. Hence, participants tend to report to platform moderators if they expect only deleting the abuser's account from the platform can prevent further dissemination of their personal information, as P3 described, *"I believe [platform moderators] are the ones that could solve this problem because I don't need any of these words to be removed and feel the particular account should be deactivated and permanently deleted from the platform."*

Time and resources. Several participants also believe that community moderators have more time and resources to review user reports than platform moderators. As a result, participants may report to community moderators if they expect more efficient moderation actions to protect their privacy. P6 observed that *"I'll say there are way more community moderators than platform moderators. And also during the time of the night and weekend hours, they would probably be able to moderate while platform moderators are out of duty."*

On the other hand, participants are also concerned that platform or community moderators might abuse the personal information in their reports, such as taking sides with the abuser, leaking information, or doxxing the reporting person. However, a majority of participants have more trust in platform moderators than community moderators to not be abusive or privacy-violating for the following reasons.

Expertise. First, almost all participants perceive platform moderators as more professional in reviewing user reports and trained on how to protect users' privacy. Therefore, platform moderators are believed to recognize the sensitivity of personal information, adhere to codes of conduct about how to access and share information, and refrain from making biased decisions from their personal standpoints. In contrast, as P6 described, *"community moderators are a person voted by the public and probably not versed in privacy and could abuse that information to track down people."*

Accountability. Second, many participants think platform moderators are more consistently accountable for potential abuse than community moderators. Platform moderators are supervised by the platform, thus there are more consequences if they violate users' trust; for example, *"getting fired or losing health insurance [P11]."* In contrast, the accountability of community moderators varies across communities. Community moderators who are invested in communities care about their reputation and *"they would have more to lose by losing their reputation and their ability to use the platform [P9],"* while community moderators in a workplace might only be loyal to *"human resources or the management rather than the users and the community itself [P9]."*

Personal connections. Moreover, participants also believe that community moderators are more likely to have personal connections with the reporting or the reported person and have motivations to misuse personal information in the reports. P6 gave an example: *"if [community moderators] don't like someone in the chatroom, might do harm with the information with the advantage they possess over some people."* In comparison, platform moderators are perceived to be distant and unlikely to take a personal grudge against users.

Time and resources. While having more moderation time and resources can be a motivator for people to report to community moderators, participants also believe that it increases the chances that community moderators might notice the personal information in users' reports and take advantage of it. P4 clearly expressed his concerns: *"I'm assuming that the scope [of community moderators] is smaller and they're responsible for moderating fewer people. My concern would be that they have more time to be worried about me, or they have more time to actually think about this one report."*

4.4.3 Which information should I share?

As we have discussed in § 4.1.1, participants are more concerned about disclosing to moderators the content of their conversations than their account and device information when reporting unwanted messages. In the following, we describe how participants carefully decide which information in their conversation they want to share with moderators to both provide evidence against abusers and protect their own privacy.

Towards making an informed moderation decision. Participants are willing to share information that they believe is important for an informed and fair moderation decision. More than half of the participants chose to share the context around the reported message with moderators to both underline the abusiveness of the reported person and to show their own innocence. For example, P13 suggested that abusive messages can be contextual: *"Sometimes the hate speech can be sarcasm and mockery. It can seem unoffensive when out of context, but with the context, it might make more sense,"* whereas P4 tried to show his civility: *"The other person in this chat is definitely being abusive. But my concern is that if I don't share my own messages, then from the perspective of moderators, I could have also said something horrible."*

However, participants tend to share less information and select less sensitive information as long as they believe what they are sharing is enough for an informed and fair moderation decision. For instance, several participants choose not to share the context around the reported message because they believe the reported messages are abusive on the surface and *"the other context might not really help in this position [P3]."* When receiving abusive messages repeatedly, participants choose to only report the abusive messages that do not include their personal information. P2 decided not to report the abusive photo of her: *"Because the whole point of getting this person reported is to have the messages deleted and then banned. And based on the first two [abusive messages without photos], I think it's pretty obvious some action will be taken."*

Less willing to share identifying information. In general, participants feel less comfortable sharing with moderators messages containing personally identifying information (e.g., phone number, email address, workplace) than those containing information about personal preferences. P4 compared these two kinds of information as follows: *"I think a platform moderator is probably not interested in my personal life. But if I'm revealing where I live, contact information, or my workplace, that would concern me more."* This is because participants believe identifying information, once put together with information about personal preferences, opens the way for greater abuse, as P10 explained, *"if I make a comment about my political views and it's not attached to any identifier, then you can't trace it back to me and I'm not really concerned about that."*

As a result, if participants are certain that their reports are anonymized to platform moderators, they are willing to share more contextual information due to the belief that moderators can not easily associate their personal preferences with their identity. For example, in cases when platform moderators only have access to anonymized reports, P2 told us: *"I'm comfortable sharing the entire chat. But if they have access to my email, to my phone number, to linked social media, then I wouldn't be comfortable sharing all of it."*

On the other hand, participants are less willing to share information about personal preferences (e.g., medical history, personal photos, political views) with community moderators. In their mental models, participants perceive platform moderators as *"in a far, far away location and removed from the applications infrastructure [P13]"* but community moderators as someone they might run into in their personal life. Therefore, disclosing personal preferences to community moderators not only introduces the risk of abuse but also feels to participants more awkward and confrontational. When receiving abusive messages with his intimate photos, P11 clearly expressed his concerns about context collapse: *"The concern is that when people see the photos you change their impression in some way. [So] I would prefer platform moderators than who I*

have some relationship with, even if it's in a vague sense that they moderate the community...I'd rather it not be someone who has any ties to me at all."

5 Discussion

5.1 Trust in the reporting system

In participants' mental models of how reporting works, we observed their uncertainties and privacy concerns, finding that a considerable number of participants feel uncertain about how E2EE platforms will protect data from reports at rest against malicious parties or whether platforms will appropriate data for their unwarranted use. In the end, participants often find themselves *having to* trust these E2EE platforms to act in users' best interest. As P6 summarized concisely, *"No, there is absolutely no guarantee [E2EE platforms] can't get the information. Otherwise, you would compile your own server and own client which nobody would do. So no guarantee, only trusting what the company says and its reputation."* From the interview, we identified three primary factors that influence people's trust in E2EE messaging platforms.

- **Open source:** First, participants have more trust in open-source messaging platforms whose source codes and encryption protocols are open to external audit and review. Being open-source also means easy replication and replaceability of the original platform, rendering the platform less likely to violate users' trust.
- **Business model:** Second, people have more trust in platforms powered by donation and partnership than those powered by advertising and marketing because they believe the latter are more motivated to collect users' data and then generate user profiles [30].
- **Historical behavior:** Finally, historical behavior is another factor that influences people's trust in E2EE platforms [8, 28]. Participants keep an eye on E2EE platforms' behaviors to observe *"a point where they turn and they start acting against their users' interests [P8]."* More privacy-conscious participants further extend their observations to developers and leaders of E2EE platforms.

Prior research has suggested users' trust in certain technologies may determine whether they can fully utilize the privacy and security advantages these technologies offer [28, 39, 52]. Here we also found that users' trust in E2EE platforms and moderators significantly influences their reporting decisions. For instance, users make assessments of potential privacy risks of disclosing sensitive information to platforms based on their trust in platforms, which then factor into their decision about whether they should report. We have also observed users' varying levels of trust in platform moderators and community moderators determine to whom they would prefer to report and how much information they would share with them. Some participants may even leave a community if they think moderators are untrustworthy. These findings highlight trust as another important factor in designing a more privacy-preserving reporting system. Prior research has underscored people's mental models of privacy-preserving technologies in shaping their use of these technologies [3]. However, even if users have a functional mental model of reporting, users may still feel vulnerable because they do not trust the platforms and moderators that operate in this system. Future work should investigate how to increase users' trust in platforms and moderators through design.

5.2 Privacy calculus in user reporting

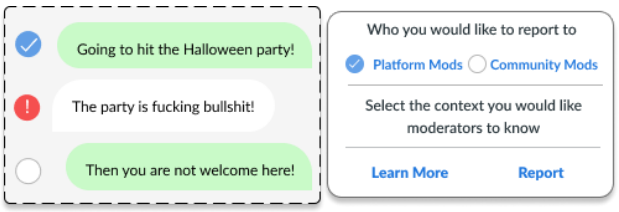
Some of our findings can be analyzed within the broader framework of privacy calculus, a cost-benefit trade-off analysis that accounts for inhibitors and drivers that simultaneously influence the decision on whether to disclose information or not [22, 23, 44]. For instance, our research revealed that individuals' decisions about whether they should report are a result of weighing the privacy risks of reporting (e.g., sharing sensitive information with platforms) against its privacy benefits (e.g., reducing further exposure of sensitive information to more people). Additionally, people make nuanced decisions about whom to report and which information to share in order to minimize the privacy risks of reporting. Similar to prior research in the context of e-commerce transactions, we also observed the impact of trust in platforms on users' evaluation of benefits and risks [22]. Future work should further explore other factors, such as individuals' propensity to trust and their control over shared information, within the context of reporting.

5.3 Design implications for user reporting on E2EE platforms

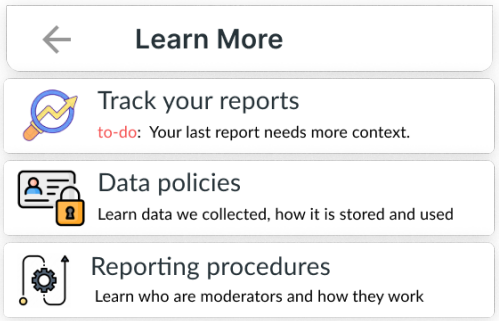
Echoing prior research that advocates for providers to focus on users' needs and experiences when building out their abuse-reporting functionality [35, 53], we further articulate three design implications for user reporting in E2EE platforms and provide a mockup of some of our proposals in Fig. 3.

5.3.1 More granularity when compiling reports

Frustrated at their exclusion from the reporting process, nearly all participants desired greater agency when compiling their reports to "control the contexts in which their information flows" [46]. First, given the contextual nature of online harassment, participants should be empowered to select which messages to be included in a report. Users could select individual messages in order to exclude messages with sensitive information from a report, or include messages that are important for moderation but outside of the immediate context. We also envision the possibility of obfuscating sensitive information in individual messages. For example, victims of non-consensual intimate imagery might desire to mask their



(a) A more granular reporting interface



(b) An interface to learn more and interact with moderators

Figure 3: **Design implications.** (a) When users choose to report a message, they can select the messages to be included in the report and to whom the report goes. (b) Users can learn more about reporting decisions for individual reports and reporting systems in general, and interact with moderators.

intimate photos but still prove to moderators the existence of these photos. Back-and-forth interactions between users and moderators might also be desirable in order to enable *progressive self-disclosure*, allowing users and moderators to reach the right balance between sharing too much and too little. For instance, users could withhold sensitive information on the first try but add additional specific context upon moderators' request.

Second, users will also benefit from the ability to choose to whom to report, as reporting to different types of moderators has different privacy implications for them. In reality, although users have the choice to report to platform moderators or community moderators, reporting to the latter typically relies on ad hoc approaches such as emails, dedicated channels, or direct messages [25]. Platforms should implement more structured reporting systems to support reporting to community moderators. There is also the possibility that community moderators are malicious or just negligent, and in such cases, the ability to escalate reports to platform moderators after attempting to report to community moderators could help hold community moderators accountable.

However, potential abusers are likely to exploit some of these features to conduct falsified abuse reports [47]. For example, they may carefully skip the context so that an innocuous joke looks like harassment. While participants consider this possibility as an inevitable sacrifice for more user agency, future work should explore how to uncover falsified abuse reports without compromising users' agency.

5.3.2 More interaction between stakeholders

While a report may involve many stakeholders, including community members, the reporting and reported person, and moderators, existing reporting systems only allow a single, one-way interaction between the reporting person and moderators. Future reporting systems should enable more flexible interactions between stakeholders in reporting procedures. For example, the different roles played by platform and community moderators highlight an opportunity for them to interact and collaborate: community moderators can request information such as metadata or cross-community reporting history from platform moderators, who could request information about the context of a report from community moderators in turn. Another idea is to enable the reported person to provide competing evidence and show that the report filed against them is misleading. Finally, other community members might be asked to corroborate a report, especially in cases where victims of abusive behaviors are too overwhelmed to report all of them [45] or an abusive message is sent in a group chat. These interactions could be one way to gather more information to uncover falsified abuse reports mentioned previously.

However, there are also various risks raised by these interactions, and we emphasize various stakeholders should only be invited after gaining explicit consent from the reporting person [33]. For instance, the reporting person may not want to have their report shared with platform/community moderators for privacy reasons. Besides, if the reported person is made aware of the report and infers the reporting person, they may choose to retaliate against them. In addition, involving malicious community members in corroborating a report may lead to the report being sabotaged. Therefore, we propose that the reporting person should have the ultimate say in weighing the benefits versus risks and be able to decide whether to invite other stakeholders into the reporting procedure.

5.3.3 Help develop proper mental models

Our findings indicate that users develop their mental models of reporting and then act accordingly. They may refrain from reporting abusive messages with their sensitive information if they do not believe reports are anonymized to moderators, or from reporting at all if they believe platforms appropriate data from reports for purposes such as advertising. A correct mental model of reporting is also essential if users are provided more agency to make granular reporting decisions, as users may have difficulty understanding the privacy implications of their choices.

These findings underscore the need to help users develop properly functional mental models. First, platforms should be

more transparent about their data policies and procedures of reporting systems. Privacy-conscious participants are eager to know what data is shared by default with the platform versus moderators, and how data is stored and used over time. More disclosure about the workflows of moderation teams and moderation statistics (e.g., how many reports are denied, how frequently each kind of moderation action is made) is also crucial for users to develop trust and confidence in the reporting system. Given users already struggle to understand E2EE itself [2, 55], E2EE platforms should provide more tutorials about their reporting systems and use less technical language.

6 Limitations and Future Work

To highlight privacy considerations, we narrowed our focus to reporting systems on E2EE platforms instead of online platforms. While future work is needed to determine how these findings generalize to non-E2EE settings, we expect some of our findings to be generalizable. For instance, we saw that participants' decisions about whether to report and to whom to report were not always motivated by their understanding of how E2EE works. This is partly because several participants who use platforms like WhatsApp and Messenger are primarily using them due to peer influence; indeed, some of them had a limited understanding of E2EE [4, 21].

There are also limitations regarding our methods. Due to our recruiting method, while we had a few participants with experience with online harassment, most participants only experienced spam on E2EE platforms. While prior research has suggested that privacy-related online harassment is increasingly pervasive [15, 59], future work should also conduct a more comprehensive survey to understand the landscape of online harassment on E2EE platforms. While we created hypothetical harassment scenarios to help participants put themselves in the shoes of the reporting person, their reporting decisions may still deviate from people who have personally experienced harassment. Future study on new reporting designs should involve more insights and feedback from people who have experienced privacy-related harassment. Given our qualitative approach and purposive sampling, a smaller sample size of 16 is suitable—however, it also means that the mental models we collected from participants do not cover all E2EE platforms and communities, which may vary greatly in terms of their technical affordances and governance models. Future work may provide a more comprehensive analysis via a larger quantitative study. Moreover, we cannot rule out the possibility that participants' privacy considerations were implicitly biased by their preconceptions due to our use of a mock interface based on WhatsApp.

Due to the scope of our research, we leave the following research questions for future work. First, while we anecdotally observed that our participants who have experienced reporting harassment could better relate to situations and advocated for greater user agency during interviews, we lacked enough data to compare them with participants without reporting experience to understand how they influenced the study results. Furthermore, our focus in this research was on examining how individuals' perceived differences between platform and community moderators influence their trust in them. However, given the diverse nature of community settings, future research should investigate the impact of various characteristics of communities on people's trust in moderators. Finally, as our study focused on gathering insights from community members, we omitted the perspective of community/platform moderators who must decide what action to take based on reports they receive. Due to limited access to platform moderators and opaque internal report handling processes at companies, we intend to address this challenge by interviewing volunteer moderators and server admins on community-operated platforms such as Matrix, Reddit, or Mastodon instead.

7 Conclusions

Prior research has advocated for user reporting as the moderation approach that most preserves the privacy guarantees of E2EE platforms. However, if users still have privacy concerns or even unfounded misgivings about reporting, user reporting loses its effectiveness in addressing online harassment. Through semi-structured interviews with E2EE users, we uncovered users' mental models and privacy concerns and considerations regarding reporting on E2EE messaging platforms. We indeed find that users have privacy concerns about reporting that sometimes lead them to refrain from reporting. Participants also have differing mental models and frequently expressed uncertainty in our interviews about aspects of how reporting works—details that are difficult for the public to validate given the lack of platform transparency. Instead, they often need to rely on their trust in platforms to weigh privacy risks and protections of reporting. Given our findings around the contextual nature of people's privacy concerns, we argue that in order for reporting systems to truly be effective, they need to provide users with a greater ability to navigate trade-offs when it comes to privacy risks.

Acknowledgments

This work was supported by the NSF SaTC award #2120497. We would like to thank Tadayoshi Kohno, members of the Social Futures Lab at the University of Washington, and collaborators at Cornell Tech for their invaluable help in this project. We also would like to thank our anonymous reviewers for their insightful feedback. Finally, we would like to express our heartfelt thanks to all the participants who dedicated their time and effort to participate in our study.

References

[1] Hal Abelson, Ross Anderson, Steven M Bellovin, Josh Benaloh, Matt Blaze, Jon Callas, Whitfield Diffie, Susan Landau, Peter G Neumann, Ronald L Rivest, et al. Bugs in our pockets: The risks of client-side scanning. *arXiv preprint arXiv:2110.07450*, 2021.

[2] Ruba Abu-Salma, Kat Krol, Simon Parkin, Victoria Koh, Kevin Kwan, Jazib Mahboob, Zahra Traboulsi, and M Angela Sasse. The security blanket of the chat world: An analytic evaluation and a user study of telegram. In *2nd European Workshop on Usable Security*. Internet Society, 2017.

[3] Ruba Abu-Salma, Elissa M Redmiles, Blase Ur, and Miranda Wei. Exploring user mental models of {End-to-End} encrypted communication tools. In *8th USENIX Workshop on Free and Open Communications on the Internet (FOCI 18)*, 2018.

[4] Ruba Abu-Salma, M Angela Sasse, Joseph Bonneau, Anastasia Danilova, Alena Naiakshina, and Matthew Smith. Obstacles to the adoption of secure communication tools. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 137–153. IEEE, 2017.

[5] Mark S Ackerman, Lorrie Faith Cranor, and Joseph Reagle. Privacy in e-commerce: examining user scenarios and privacy preferences. In *Proceedings of the 1st ACM Conference on Electronic Commerce*, pages 1–8, 1999.

[6] Adrian Chen. The laborers who keep dick pics and beheadings out of your facebook feed. "https://www.wired.com/2014/10/content-moderation/", 2014. [Online; accessed 18-Jan-2023].

[7] Zahra Ashktorab and Jessica Vitak. Designing cyberbullying mitigation and prevention solutions through participatory design with teenagers. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 3895–3905, 2016.

[8] Erinn Atwater, Cecylia Bocovich, Urs Hengartner, Ed Lank, and Ian Goldberg. Leading johnny to water: Designing for usability and trust. In *Proceedings of the Eleventh USENIX Conference on Usable Privacy and Security*, pages 69–88, 2015.

[9] Katie Benner and Mike Isaac. Child-welfare activists attack facebook over encryption plans. The New York Times, 2022.

[10] Monika Bickert. Publishing our internal enforcement guidelines and expanding our appeals process, April 2018.

[11] Lukas Bieringer, Kathrin Grosse, Michael Backes, Battista Biggio, and Katharina Krombholz. Industrial practitioners' mental models of adversarial machine learning. In *Eighteenth Symposium on Usable Privacy and Security (SOUPS 2022)*, pages 97–116, 2022.

[12] Jeremy Blackburn and Haewoon Kwak. Stfu noob! predicting crowdsourced decisions on toxic behavior in online games. In *Proceedings of the 23rd international conference on World wide web*, pages 877–888, 2014.

[13] Virginia Braun, Victoria Clarke, Nikki Hayfield, and Gareth Terry. Thematic analysis. In Pranee Liamputtong, editor, *Handbook of Research Methods in Health Social Sciences*, pages 843–860. Springer Singapore, Singapore, 2019.

[14] Facebook Help Center. End-to-end encryption.

[15] Pew Research Center. Nearly half of those who have been harassed online know their harasser, 2017.

[16] Pew Research Center. The state of online harassment, 2021.

[17] Eshwar Chandrasekharan, Chaitrali Gandhi, Matthew Wortley Mustelier, and Eric Gilbert. Crossmod: A Cross-Community Learning-based System to Assist Reddit Moderators. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–30, nov 2019.

[18] Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. Anyone can become a troll: Causes of trolling behavior in online discussions. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, pages 1217–1230, 2017.

[19] Kate Crawford and Tarleton Gillespie. What is a flag for? social media reporting tools and the vocabulary of complaint. *New Media & Society*, 18(3):410–428, 2016.

[20] Rachel Cummings, Gabriel Kaptchuk, and Elissa M Redmiles. " i need a better description": An investigation into user expectations for differential privacy. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, pages 3037–3052, 2021.

[21] Alexander De Luca, Sauvik Das, Martin Ortlieb, Iulia Ion, and Ben Laurie. Expert and {Non-Expert} attitudes towards (secure) instant messaging. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*, pages 147–157, 2016.

[22] Tamara Dinev, Massimo Bellotto, Paul Hart, Vincenzo Russo, Ilaria Serra, and Christian Colautti. Privacy calculus model in e-commerce–a study of italy and the united states. *European Journal of Information Systems*, 15(4):389–402, 2006.

[23] Tamara Dinev and Paul Hart. An extended privacy calculus model for e-commerce transactions. *Information systems research*, 17(1):61–80, 2006.

[24] Discord. What is discord. https://discord.com/safety/360044149331-what-is-discord. [Online; accessed 28-Jan-2023].

[25] Discord. Best practices for reporting tools, 2022. [Online; accessed 28-Jan-2023].

[26] Yevgeniy Dodis, Paul Grubbs, Thomas Ristenpart, and Joanne Woodage. Fast message franking: From invisible salamanders to encryptment. In *Annual International Cryptology Conference*, pages 155–186. Springer, 2018.

[27] Facebook. Facebook: Messenger secret conversations technical whitepaper, 2017.

[28] Sascha Fahl, Marian Harbach, Thomas Muders, Matthew Smith, and Uwe Sander. Helping johnny 2.0 to encrypt his facebook conversations. In *Proceedings of the eighth symposium on usable privacy and security*, pages 1–17, 2012.

[29] Brown Farinholt, Mohammad Rezaeirad, Paul Pearce, Hitesh Dharmdasani, Haikuo Yin, Stevens Le Blond, Damon McCoy, and Kirill Levchenko. To catch a ratter: Monitoring the behavior of amateur darkcomet rat operators in the wild. In *2017 IEEE symposium on security and privacy (SP)*, pages 770–787. Ieee, 2017.

[30] Nina Gerber, Verena Zimmermann, Birgit Henhapl, Sinem Emeröz, and Melanie Volkamer. Finally johnny can encrypt: But does this make him feel more secure? In *Proceedings of the 13th International Conference on Availability, Reliability and Security*, pages 1–10, 2018.

[31] Sarah A Gilbert. Towards intersectional moderation: An alternative model of moderation built on care and power. *arXiv preprint arXiv:2305.11250*, 2023.

[32] Google. Youtube community guidelines enforcement in google's tranparency report for 2018., 2018.

[33] Jane Im, Jill Dimond, Melody Berton, Una Lee, Katherine Mustelier, Mark S Ackerman, and Eric Gilbert. Yes: Affirmative consent as a theoretical framework for understanding and imagining social platforms. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2021.

[34] Adrianne Jeffries. Meet 'swatting,' the dangerous prank that could get someone killed, 2013. [Online; accessed 28-Jan-2023].

[35] Seny Kamara, Mallory Knodel, Emma Llansó, Greg Nojeim, Lucy Qin, Dhanaraj Thakur, and Caitlin Vogus. Outside looking in: Approaches to content moderation in end-to-end encrypted systems. *arXiv preprint arXiv:2202.04617*, 2022.

[36] Ruogu Kang, Laura Dabbish, Nathaniel Fruchter, and Sara Kiesler. {"My} data just goes {Everywhere:"} user mental models of the internet and implications for privacy and security. In *Eleventh Symposium on Usable Privacy and Security (SOUPS 2015)*, pages 39–52, 2015.

[37] Predrag Klasnja, Sunny Consolvo, Jaeyeon Jung, Benjamin M Greenstein, Louis LeGrand, Pauline Powledge, and David Wetherall. " when i am on wi-fi, i am fearless" privacy concerns & practices in eeryday wi-fi use. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1993–2002, 2009.

[38] Yubo Kou and Xinning Gui. Flag and flaggability in automated moderation: The case of reporting toxic behavior in an online game community. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2021.

[39] Katharina Krombholz, Karoline Busse, Katharina Pfeffer, Matthew Smith, and Emanuel Von Zezschwitz. " if https were secure, i wouldn't need 2fa"-end user and administrator mental models of https. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 246–263. IEEE, 2019.

[40] Cliff Lampe and Paul Resnick. Slash (dot) and burn: distributed moderation in a large online conversation space. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 543–550, 2004.

[41] Anti-Defamation League. Online hate and harassment: The american experience 2022, 2022.

[42] Amanda Lenhart, Michele Ybarra, and Myeshia Price-Feeney. Nonconsensual image sharing: one in 25 americans has been a victim of "revenge porn", 2016.

[43] Ada Lerner, Eric Zeng, and Franziska Roesner. Confidante: Usable encrypted email: A case study with lawyers and journalists. In *2017 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 385–400. IEEE, 2017.

[44] Han Li, Rathindra Sarathy, and Heng Xu. Understanding situational online information disclosure as a privacy calculus. *Journal of Computer Information Systems*, 51(1):62–71, 2010.

[45] Kaitlin Mahar, Amy X Zhang, and David Karger. Squadbox: A tool to combat email harassment using friend-sourced moderation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2018.

[46] Alice E Marwick and Danah Boyd. Networked privacy: How teenagers negotiate context in social media. *New media & society*, 16(7):1051–1067, 2014.

[47] J Nathan Matias, Amy Johnson, Whitney Erin Boesel, Brian Keegan, Jaclyn Friedman, and Charlie DeTar. Reporting, reviewing, and responding to harassment on twitter. *arXiv preprint arXiv:1505.03359*, 2015.

[48] Matrix. Introduction to matrix. https://matrix.org/docs/guides/introduction. [Online; accessed 28-Jan-2023].

[49] Matrix. Moderation in matrix. https://matrix.org/docs/guides/moderation#reporting-bad-content, 2022. [Online; accessed 28-Jan-2023].

[50] Susan E McGregor, Polina Charters, Tobin Holliday, and Franziska Roesner. Investigating the computer security practices and needs of journalists. In *24th USENIX Security Symposium (USENIX Security 15)*, pages 399–414, 2015.

[51] George R Milne, George Pettinico, Fatima M Hajjat, and Ereni Markos. Information sensitivity typology: Mapping the degree and type of risk consumers perceive in personal data sharing. *Journal of Consumer Affairs*, 51(1):133–161, 2017.

[52] Alena Naiakshina, Anastasia Danilova, Sergej Dechand, Kat Krol, M Angela Sasse, and Matthew Smith. Poster: Mental models-user understanding of messaging and encryption. In *Proceedings of European Symposium on Security and Privacy. http://www. ieee-security. org/TC/EuroSP2016/posters/number18. pdf*, 2016.

[53] Riana Pfefferkorn. Content-oblivious trust and safety techniques: Results from a survey of online service providers. *Journal of Online Trust and Safety*, 1(2), 2022.

[54] Sarah T Roberts. Commercial content moderation: Digital laborers' dirty work. In *The Intersectional Internet: Race, Sex, Class and Culture Online*. Peter Lang Publishing, 2016.

[55] Svenja Schröder, Markus Huber, David Wind, and Christoph Rottermanner. When signal hits the fan: On the usability and security of state-of-the-art secure mobile messaging. In *European Workshop on Usable Security. IEEE*, pages 1–7, 2016.

[56] Julia Sinclair-Palm and Kit Chokly. 'it'sa giant faux pas': exploring young trans people's beliefs about deadnaming and the term deadname. *Journal of LGBT Youth*, pages 1–20, 2022.

[57] Slack. Join a slack workspace. https://slack.com/help/articles/212675257-Join-a-Slack-workspace. [Online; accessed 28-Jan-2023].

[58] Peter Snyder, Periwinkle Doerfler, Chris Kanich, and Damon McCoy. Fifteen minutes of unwanted fame: Detecting and characterizing doxing. In *proceedings of the 2017 Internet Measurement Conference*, pages 432–444, 2017.

[59] Kurt Thomas, Devdatta Akhawe, Michael Bailey, Dan Boneh, Elie Bursztein, Sunny Consolvo, Nicola Dell, Zakir Durumeric, Patrick Gage Kelley, Deepak Kumar, et al. Sok: Hate, harassment, and the changing landscape of online abuse. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 247–267. IEEE, 2021.

[60] Nirvan Tyagi, Paul Grubbs, Julia Len, Ian Miers, and Thomas Ristenpart. Asymmetric message franking: Content moderation for metadata-private end-to-end encryption. In *Advances in Cryptology–CRYPTO 2019: 39th Annual International Cryptology Conference, Santa Barbara, CA, USA, August 18–22, 2019, Proceedings, Part III 39*, pages 222–250. Springer, 2019.

[61] Kathleen Van Royen, Karolien Poels, and Heidi Vandebosch. Help, i am losing control! examining the reporting of sexual harassment by adolescents to social networking sites. *Cyberpsychology, Behavior, and Social Networking*, 19(1):16–22, 2016.

[62] Bertie Vidgen and Leon Derczynski. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *Plos one*, 15(12):e0243300, 2020.

[63] WhatsApp. Communities now available! "https://blog.whatsapp.com/communities-now-available", 2022. Online; accessed 28-Jan-2023.

[64] WhatsApp. How to block and report contacts. "https://faq.whatsapp.com/1142481766359885/?cms_platform=android", 2022. Online; accessed 28-Jan-2023.

[65] Andy Zhao and Zhaodi Chen. Let's report our rivals: how chinese fandoms game content moderation to restrain opposing voices. *Journal of Quantitative Description: Digital Media*, 3, 2023.

A Appendix

A.1 Interview Protocol

Thank you for taking the time to participate in our interview today. We appreciate your help with this research study. The interview will be about an hour. During the interview, we will ask you about your experience with and thoughts about reporting on encrypted messaging platforms. We will also show you several hypothetical scenarios and ask about your opinions about reporting and related privacy concerns. Please feel free to skip questions or pause the interview if at any point you feel uncomfortable answering the questions.

Background: understand the use of E2EE platforms

- Why did you start using E2EE messaging platforms? To what extent was privacy protection your motivation for using E2EE platforms?
- Who do you usually chat with on each of these encrypted platforms you use?
- Who do you think can have access to your messages on E2EE platforms?
- Have you ever tried to report an account, a message, or a conversation before on these platforms? Do you feel comfortable explaining the context of your reporting? Feel free to skip this question.
- Have you ever received or witnessed unwanted messages such as bullying, hate speech, and harassment before on these platforms? Do you feel comfortable describing your experience? Feel free to skip this question.

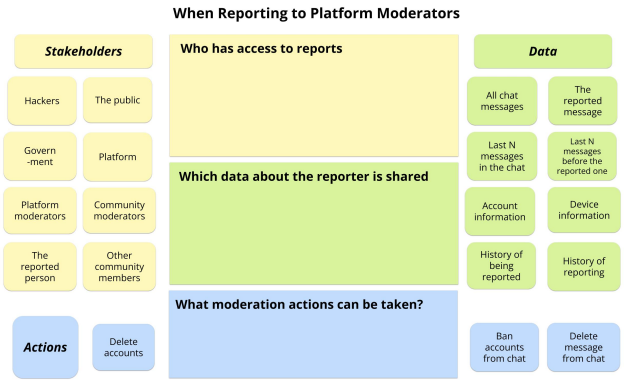


Figure 4: **Interactive digital board that allow users to explain their mental models of reporting on E2EE.** Participants are encouraged to drag labeled cards to indicate their mental models.

Section I: eliciting mental models. For reporting to platform moderators and to community moderators, ask the following questions respectively. During the interview, ask participants to draw the information flow of reports and help them refine the flow by the following interview questions.

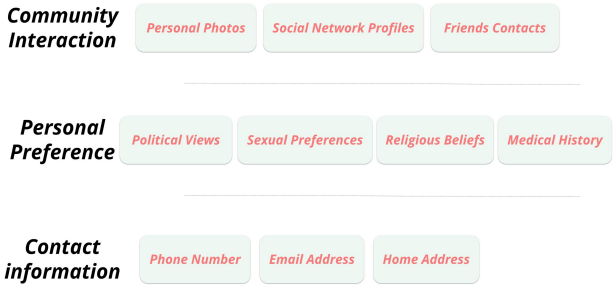


Figure 5: **Harassment scenario options.** During the interview, participants were encouraged to select the items that they relate to and are comfortable discussing from this board listing all the possible options in a hypothetical harassment scenario. Abusive language is masked by default to protect participants from unnecessary harm.

- Who do you think can have access to your reports?
- Which information do you think is shared with moderators when you make a report?
- What actions do you think moderators can take for your reports?
- Are you worried about these data will be shared with each stakeholder here?
- Do you imagine that these shared data will still be used elsewhere after report decisions are made?
- Have you noticed or experienced falsified abuse reports?

Section II: ask privacy concerns about reporting. First, ask participants to choose the category that they share frequently and consider sensitive but also feel comfortable talking about. In each selected scenario, ask the following questions with the introduction: Imagine you are using an E2EE group messaging platform like WhatsApp or Signal. You received a few abusive messages from another chat member.

- Will you choose to report these messages or not?
- Who are you going to report to, community moderators or platform moderators?
- What actions do you think moderators can take for your reports?
- If you are allowed to choose which part of messages is filed in a report, what messages would you like to choose for a report?
- If you are allowed to choose which part of metadata is filed in a report, which metadata information about you and the reported person do you would like to choose for a report?

A.2 Hypothetical Harassment Scenarios

Content Warning: *The following scenarios contain hate speech and offensive language.*

Political views
A: Many of the families who are being separated at the border and who are being kept in poor conditions did try to cross legally, through legal ports of entry seeking asylum.
B: They are their own country's problems. Not the problems of the United States. They are fucking invaders. [description: *B expressed hate speech about illegal immigrants*]
A: Calm down, you aren't even trying to be civil.
B: Then deport you big trash and your friends to Mexico. If you want to help the immigration problem, start there! Problem solved for all, you pussy liberals [description: *B expressed hate speech about people's empathy towards illegal immigrants*]

Sexual preference
A: Hey everyone, I just want to share that I am a woman and go by Samantha and use she/her pronouns.
B: Ugh, that's fucking disgusting, you are a boy, you have a penis, and you are Robert. you cannot just change your name and gender. [description: *B: Insults to you about your being transgender*]
A: I had hoped that you'd be more supportive.
B: This is so sad and pathetic, you are such a loser and a sissy, you are dead to me. And I will tell everyone about this. [description: *B expressed insults to you about your being transgender and threatens to tell everyone*]

Religious belief
A: As a Muslim girl, am I welcome here?!
B: No, you are not! We did NOT invite Muslims, Africans, and all sorts here! Fuck that Muslim piece of shit. [description: *B expressed hate speech towards Muslim people*]
A: I don't believe in violence and we should treat each other with love.
B: As Muslims, go back to your Muslim hell hole countries, you ungrateful B****! [description: *B expressed insults to you about your being Muslims*]

Personal photos
A: Just getting some cake from the Cafe down the street. (a photo of yourself holding a cake)
B: You are so fat, why do you eat so much? [description: *B expressed insults to you about your body*]
A: Excuse me?
B: I am going to re-share this edited photo to Facebook (a photo of A holding a cake; now with the caption "fat cat, fatty cat") [description: *B expressed insults to you about your body and shared an edited but now abusive photo of you holding the cake*]

Phone number
(Context: A shared their phone number with B when they were on a company vacation together, the internet was spotty, and they needed to use SMS to communicate.)

A: I am planning on visiting my family in Turkey during this holiday, and may be unavailable for the next few days.
B: Ugh, I didn't know that you were an immigrant. Immigrants are smelly, shitty, and taking over our jobs. [description: *B expressed hate speech towards immigrants*]
A: Wait, what?
B: Get back to your country. I am going to sign up for Tinder with your phone number +1 2045661223 and swipe right on every guy. [description: *B expressed hate speech towards immigrants and threats to overwhelm you with spam and stalkers using your phone number*]

Email address
(Context: A shared their work email address with B for an office meeting.)

A: I am planning on visiting my family in Turkey during this holiday, and may be unavailable for the next few days.
B: Ugh, I didn't know that you were an immigrant. Immigrants are smelly, shitty, and taking over our jobs. [description: *B expressed hate speech towards immigrants*]
A: Wait, what?
B: Get back to your country. I am going to sign up for gay twink porn with your work email address alice@bigcompany.co [description: *B expressed hate speech towards immigrants and threats to overwhelm you with spam and stalkers using your email address*]

Home address
(Context: A shared their home address with B for the office secret Santa list.)

A: I am planning on visiting my family in Turkey during this holiday, and may be unavailable for the next few days.
B: Ugh, I didn't know that you were an immigrant. Immigrants are smelly, shitty, and taking over our jobs. [description: *B expressed hate speech towards immigrants*]
A: Wait, what?
B: Get back to your country. I know that your address is 4200 11th Ave NE, XXX. I am going to call the police to your address to forcefully kick you out. [description: *B expressed hate speech towards immigrants and threats to overwhelm you with false police calls using your home address*]

Evaluating User Behavior in Smartphone Security: A Psychometric Perspective

Hsiao-Ying Huang[1], Soteris Demetriou [2], Muhammad Hassan[1], Güliz Seray Tuncay [3], Carl A. Gunter [1], and Masooda Bashir[1]

[1]University of Illinois at Urbana-Champaign, {*hhuang65*[*], *mhassa42, mnb, cgunter*}@illinois.edu
[2]Imperial College London, *s.demetriou@imperial.ac.uk*
[3]Google, *gulizseray@google.com*

Abstract

Smartphones have become an essential part of our modern society. Their popularity and ever-increasing relevance in our daily lives make these devices an integral part of our computing ecosystem. Yet, we know little about smartphone users and their security behaviors. In this paper, we report our development and testing of a new 14-item Smartphone Security Behavioral Scale (SSBS) which provides a measurement of users' smartphone security behavior considering both technical and social strategies. For example, a technical strategy would be resetting the advertising ID while a social strategy would be downloading mobile applications only from an official source.The initial analysis of two-component behavioral model, based on technical versus social protection strategies, demonstrates high reliability and good fit for the social component of the behavioral scale. The technical component of the scale, which has theoretical significance, shows a marginal fit and could benefit from further improvement. This newly developed measure of smartphone security behavior is inspired by the theory of planned behavior and draws inspiration from a well-known scale of cybersecurity behavioral intention, the Security Behavior Intention Scale (SeBIS). The psychometrics of SSBS were established by surveying 1011 participants. We believe SSBS measures can enhance the understanding of human security behavior for both security researchers and HCI designers.

[*]Hsiao-Ying Huang is best reachable at hhsiaoying@gmail.com

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.
USENIX Symposium on Usable Privacy and Security (SOUPS) 2023.
August 6–8, 2023, ANAHEIM, CA., USA.

1 Introduction

Smartphones have become an essential part of modern society. In 2021, about 85% of the adults in the U.S. owned smartphones, up from just 35% in 2011 [8]. Internationally, the number of global smartphones users is estimated at 6.8 billion, marking an 86.5% increase from 2016 [50]. Smartphones are now involved in almost any daily activity watched videos, and 45% did online shopping [16]. As smartphones have become a hub for storing and accessing personal sensitive information [35], an increasing number and a diverse set of malicious parties have sought to exploit security vulnerabilities of smartphones and their users.

Mobile operating system developers have consequently been dedicated to equipping their systems with numerous counter measures (e.g., discretionary and mandatory access control, trusted computing *etc.*). However, the security of such systems still heavily relies on the behavior and decision-making of users. For instance, Android and iOS feature a permission model to enable users to decide if they want to grant mobile applications access to sensitive system resources and information. Some repackaged malware apps aim to trick users by mimicking the look-and-feel of popular legitimate apps with subtle differences in their title or logo to attract users to download, trust, and grant them permissions [66]. Other attacks target the intricate configuration properties of smartphones: attackers can exploit the vulnerabilities in the permission models to elevate their privileges and obtain unauthorized access to sensitive user data [40, 60, 65]; attackers can extract users' passwords when they access sensitive web domains (*e.g.* their bank account) from their smartphones on a public network [39]. Users who have never reset their advertising ID, can be subjected to fine-grained profiling by advertising libraries [57] [63]; users who are not attentive to the information provided by websites or applications can become the victims of phishing attacks [5, 26, 45, 61]. Since users play a such critical role in smartphone security, a better understanding of their behavior is crucial to help drive the design of better security mechanisms in mobile apps as well

as in mobile operating systems.

When it comes to user behavior on smartphones, previous studies have investigated users' perceptions, attitudes, and behavior toward smartphone security. They have found that users tend to ignore warnings [6, 23, 36], have misconceptions about the operation of smartphone security features [27, 36, 46, 56], and show minimal attempts to protect their smartphones [15, 46]. Users' careless behavior on smartphones can particularly result from their misunderstanding of the capabilities of these devices. Most smartphone users view their smartphone as just a mobile device for entertainment and communication; they are not aware it is in fact a hand-held computer vulnerable to a wide range of cyber-attacks [38] . Users are highly likely to thus address security differently on smartphones than on other devices such as laptops or PCs. In this paper, we explore a system that can can measure users' smartphone security behavior in a systematic way across contexts.

Prior studies tried to operationalize security-related concepts in different ways. Some studies adopted field observation while others employed a self-reported approach [15, 22, 34, 59]. Since field observations usually require more resources and have limitations in assessing all aspects of security behavior, most studies utilized self-reported measurements. In terms of self-reported measurements, we found many studies developed their own measurements based on computer security or adopted those from smartphone measurements in other contexts. *Therefore, based on the current literature, there is a need for a measurement system for security behavior that is standardized and specific to smartphones.*

In order to fill this research gap, we made the first step towards providing a model for measuring human *smartphone* security behavior. We grounded it on the theory of reasoned action (TRA [24]) and the related theory of planned behavior (TPB) [1]. TRA is a well-established framework for conceptualizing and explaining human behavior with widespread applications. It posits that *behavior intentions* (BI) are immediate antecedents to *behavior*. Other established self-reported measures have accordingly been constructed to attempt to measure behavior, including relevant scales developed for computer security behavior intentions [20] and attitudes [22].

Similarly, we focused on developing a necessary measurement tool for recording smartphone-specific security behavior intentions. In particular, we developed a model and conducted an analysis to support a standardized scale for measuring users' smartphone security behavior intentions (BIs) to follow expert recommendations, based on a systematic psychometric approach [47]. We present a study with two phases evaluated on a total of 1011 participants. In our phase-1 study wherein we examined if the model of general computer security BIs could be applied to smartphone security BIs. We adopted four dimensions from a well-established measurement, the Security Behavior Intention Scale (SeBIS) developed by Egelman and Peer [20] and examined the fitness of these dimensions on users' smartphone security behavior by factor analysis. Our findings indicate *smartphone security BIs entail new dimensions that are different from the model of general computer security BIs*. Therefore, in our phase-2 study, we *created a new scale measurement for smartphone security using a systematic scale development procedure*. We operationalized expert-provided guidelines for securing mobile devices by aiming to capture users' intention to comply with such guidelines in a measurement. To assess if our new measurement reliably captures such intentions that represent smartphone security behavioral constructs we are interested in, we evaluated its dimensionality, scale reliability, and construct validity. Our results show that the new scale exhibits satisfactory psychometric properties: the full scale and both of its subscales have high internal consistency, all items map uniquely on one single component, and no correlation exists between the subscales. Lastly, convergent validity is also established between the new scale and SeBIS.

Our contributions are summarized as follows:

- We found that new dimensions are important when measuring smartphone security behavior. These are different from the general security behavior model.
- We introduce a new standardized scale tailored for smartphone security behavior intentions (SSBS) which is based on two factors (i.e., technical versus social) and showed good psychometric properties with high internal consistency.

The rest of the paper is organized as follows: in related work, we reviewed the literature that are most related to our study and proposed research questions based on research gap (Section 2). We then illustrated our psychometric approach and methodology (Section 3) and described the design and results of phase-1 (Section 4) and phase 2 (Section 5) studies respectively. Lastly, in Section 6 we discuss limitations and future work and conclude the paper in Section 7.

2 Background and Related Work

Theory Background. Researchers at the intersection of computer security and social sciences have grounded their analysis (e.g. the Technology Acceptance Model – TAM [17, 18]) on end-users' usage of technology based on psychological frameworks such as the theory of reasoned action (TRA [24]) and the related theory of planned behavior (TPB) [1, 42]. These posit that behavior is immediately preceded by a behavior intention (BI). In turn, behavior intention is a function of behavioral and normative beliefs, and the behavioral beliefs are determined by an individual's attitude toward performing the behavior. Understanding users' attitudes and intentions is fundamentally conducive to plausible interpretations of end-users' behavior and factors that might affect them. Researchers have only recently tried to operationalize such concepts in computer security. In particular, Faklaris et al. developed a new self-report measure (SA-6) for quantify-

ing end-user security attitudes [22], while Egelman et al. developed a 16-item self-report Security Behavior Intentions Scale (SeBIS) that they evaluated for both internal [20] and external validity [19]. However, these measures target general computer security behavior. Given the widespread use of smartphones, there is a need for complementary measurement standards targeted specifically at *smartphone* security behavior.

Smartphone Security. Prior works have examined smartphone users' behavior. Their findings can be categorized into three realms: the inattentiveness toward security warnings and messages [23, 36], the misconceptions of smartphone security [9, 36, 46, 56], and the low level of concern for smartphone security behavior [15, 37, 46]. In terms of behavioral measurements, previous studies assessed users' smartphone security behavior through field observations and self-reported measurements. For field observations, most studies focus on two aspects: users' authentication and locking behavior on smartphones [28, 29, 31, 32] and users' behavior on granting access [4, 25, 64]. Although field observation can probe into users' actual behavior in the real world, it usually focuses on a single aspect of the behavior, making it difficult to conveniently gain a comprehensive understanding on user behavior in a short period of time. Therefore, many studies adopt a self-reported approach to measure users' smartphone security behavior.

Low Level of Concern for Smartphone Security Behavior: Research has revealed that users in general exhibit a low level of behavioral security tendency on smartphones even though they perceive certain security threats (e.g., malware, data leakage) [15]. Furthermore, even though smartphone usage has increased and technology has advanced, general security awareness remains fairly low [37]. For instance, when selecting an application, most users did not pay attention to its information on security and privacy [46], even despite having the required knowledge and understanding [2].

Only a minority of users were interested in security and agreement information and they were more security and tech-savvy [46]. This suggests that smartphone security behavior is influenced by individual factors (e.g., knowledge, personal interests, and personalities) and can vary significantly between users.

Adapting General Self-Reports to Specific Domains. There are various means to measure self-reported smartphone security behavior. The most commonly used approach in prior research has been to develop measurements by adapting more general computer security assessments or by modifying a developed measurement from previous studies. For example, Das and Khan [15] generated a 6-item measure that was adapted from Microsoft's computing safety index. Jones and Chin [34] performed a survey study to investigate students' usage and security behavior on smartphones by asking seven questions about security practices. A more recent study by Thompson *et al.* [59] designed a five-item measure to assess smartphone security behavior in a personal context, which was adapted from a security behavioral assessment in personal computer usage by Liang and Xue [41]. Another recent (2018) survey study by Verkijika [62] examined South African users' smartphone security practices by using five questions that were adapted from the measurement developed by Thompson et al. [59].

While reviewing developed security measurements (see Table 6), we found there is no standardized and targeted way to measure smartphone security behavior intentions across different contexts. Existing methods are all adopted or adapted from general computer security behavior measurement tools. However, it is possible that users' smartphone behavior can deviate from their computer behavior. For instance, Chin et al. [9] found participants' behavior and activities on smartphones were quite different from their use of laptops. For example, users were less likely to purchase and perform sensitive tasks on their smartphones because of security concerns regarding mobile devices. Moreover, none of these smartphone security measures were grounded on psychological principles that can help us better interpret and compare results.

Conclusion and Main Objective. We thus identified two key gaps in the current literature: 1) there is no standardized measurement of smartphone security behavior intentions across contexts; 2) it remains unclear if *general* computer security behavior intentions can be applied to assess *smartphone* security behavior intentions. A key goal of this study was to develop the first standardized, valid, and specialized measurement of smartphone security behavior intentions that can be used in different contexts and form the basis for studying smartphone security behavior. Toward this goal, we posed the following concrete research questions:

- **RQ1:** How adequate is the adaptation of general computer security BIs measurement to smartphone security?
- **RQ2:** If this adaptation is not adequate, can we develop a measure to capture smartphone security BIs?

3 A Psychometrics Approach

To answer these research questions, we adopted a psychometric approach. Psychometrics is a scientific approach of quantifying human psychological attributes such as personality traits, cognitive abilities, and social attitudes [44]. Well developed and widely-used security-related psychometric measurements are the "self-report measure of security attitudes" (SA-6) developed by Faklaris et al. [22] and the "Security Behavior Intentions Scale" (SeBIS) developed by Egelman and Peer [20], which are both grounded on the "Theory of Reasoned Action". They conceptualize users' *general* security behavior as a psychological construct instead of an actual

behavior. We followed the same approach to conceptualize users' *smartphone* security behavior intentions.

Evaluation Properties. We developed the Smartphone Security Behavioral Scale (SSBS), a new measure for assessing users' behavior intentions to comply with good smartphone security practices. When developing a new scale, it is important to evaluate three psychometric properties of the measurement: dimensionality, scale reliability, and convergent validity [47].

Dimensionality. Identifying dimensionality of a construct is a critical part of scale development because whether the construct is uni-dimensional or multi-dimensional will affect the structure and computing approach of scale [47]. There are two statistical approaches to determine dimensionality based on the use case. If the goal of testing is to 'explore' the unknown dimensions of a construct, the Exploratory Factor Analysis (EFA) is an appropriate method to use. If the goal is to 'confirm' or examine the existing dimensions of a construct, then the Confirmatory Factor Analysis (CFA) is a standardized way to test the fitness of the model.

Scale Reliability. In psychometrics, reliability represents the consistency of a measurement, which can be evaluated in various ways [47]. In this study, we focused on assessing "internal consistency" of the scale to determine if multiple items in a scale measure the same construct by examining Cronbach's alpha [13]. Cronbach's alpha is the mean of all possible coefficients among items [12]. The cut-off point of Cronbach's alpha is 0.70 [48] and refers to the acceptable internal consistency of the scale. In addition, considering the numbers of items can affect the score of Cronbach's alpha [12, 58], we also reported the mean of inter-item correlation (ITC), which is the average pairwise correlation among all items and provides a direct indicator of homogeneity [11].

Construct Validity. Construct validity refers to the degree to which a measurement truly reflects the concept being examined [7]. One approach is to evaluate convergent validity between the newly-developed scale and an existing scale measuring the same construct [47, 49]. Convergent validity is measured by correlational coefficients between the new measure and an existing measure. In our study, we evaluated the convergent validity between our scale and SeBIS [20] and tested if our scale measures similar constructs of security behavior.

Since there has been a well-established computer security behavioral intentions scale (SeBIS), our first step was to examine if the dimensionality of SeBIS could be applied to users' smartphone security behavior intentions. Our findings indicate different dimensions of smartphone security. Therefore, we followed a standardized procedure of scale development proposed by Netemeyer [47]. Our procedure of scale development is summarized as follows:

1. Testing the fitness of dimensional model of SeBIS on smartphone security behavior by applying CFA.
2. Defining the construct that the scale attempted to measure and generating a list of candidate questions.
3. Extracting the dimensional components of the scale by performing EFA and reducing the set of items.
4. Finalizing the scale by conducting CFA to confirm the fitness of the new scale to the intended factorial model.

Methodology. The goal of this study is to develop a measurement to assess users' security behavior intentions to comply with smartphone security advice recommended by security professionals. We conducted a two-phase online survey study, approved by our Institutional Review Board. In phase-1, we tested the four dimensions used in SeBIS [20]. Our results suggest the possibility of improving on the four dimensions of SeBIS when specializing them to smartphone security behavior intentions. In other words, users' smartphone security behavior intentions could be different from their general computer security behavior intentions. We therefore conducted a phase-2 study to develop a new measurement for smartphone security behavior intentions.

For both phases, we recruited participants from the United States via Amazon Mechanical Turk (MTurk). To ensure data quality, we integrated attention-check questions in each section of the survey. The attention-check questions were randomly inserted in the questionnaire and had similar format to other questions. Participants were required to select the choice required in the statement (for instance, *I go to grocery shopping on every Thursday. Please select 'Never'*). We removed the responses from participants who failed to correctly answer attention-check questions. We next describe the details of study design and results for each phase of the study.

4 Phase-1: Building the scale upon SeBIS

4.1 Survey design and item generation

We first developed a measurement, which we call *smartphone-SeBIS*, based on the four dimensions of SeBIS: device securement, password management, proactive awareness, and update. We generated items by revising each question in SeBIS for a smartphone context. For example, we changed the wording of questions from 'computer' to 'smartphone'. However, we encountered two challenges when using this approach. First, we found that certain questions could not be readily applied to smartphones. Secondly, certain common smartphone-specific security features were not included in SeBIS, such as biometrics, usage of applications, and app permissions. To capture a more comprehensive view of users' smartphone security behavior, we recruited security experts who independently went over each item of the first version of smartphone-SeBIS and considered how to revise old items and add new items to the survey. Overall, we had four types of item modifi-

cations: word/phrase substitution, word/phrase revision, item deletion, and item addition.

Word/Phrase Substitution. we substituted words indicating the context of a laptop or desktop machine to specifically describe a smartphone. For instance, to capture the same behavior on a smartphone device, we substituted the word "smartphone" for "laptop or tablet" in the item "*I use a password/passcode to unlock my laptop or tablet.*"

Word/Phrase Revision. some items could not be made smartphone-specific with simple substitutions. For example, "*I do not change my passwords, unless I have to*". This was revised to the following: "*I regularly change my password for online services/accounts using my smartphone,*" where we specified the password target to avoid confusion and turned the negative statement into a positive statement. We did this since participants might be biased toward taking a defensive stance against the negative behavior.

Item Deletion. Some of the SeBIS items are not applicable to the smartphone context. For instance, the item "*When browsing websites, I mouse-over links to see where they go, before clicking them*" is not applicable on mobile devices since the pointing mechanism on smartphones is different (mouse or trackpad for desktops/laptops vs finger or stylus on mobile devices). Such items were removed from the survey.

Item Addition. Several important smartphone security behaviors were not specified or included in SeBIS. For instance, significant security mechanisms introduced by Original Equipment Manufacturers (OEMs), or by the research community, become obsolete if the user roots (or jailbreaks) their smartphone. This is an important "device securement" measurement to take. Moreover, on smartphones, user privacy is preserved through a permission system that allows users to determine what device and personal information each installed third-party app can access. This mechanism can also be compromised if users become inattentive to permission requests or if they never revoke permissions from apps [23, 64]. To address such phenomena, we added relevant smartphone-specific items into the survey.

As a result of this exercise, we developed the Smartphone-SeBIS, a comprehensive instrument consisting of 20 items targeting smartphone security behaviors (see Table 5). We administered the Smartphone-SeBIS through an online survey using Amazon MTurk, where participants were asked to respond on a 5-point Likert-type scale ranging from Never' to Always'. To mitigate the potential priming effect of social desirability, we advertised our study as an investigation into 'the use of smartphones and mental health wellness' and included several related questions in the survey questionnaire. To control for potential order effects, the survey sections were randomized. After completing the questionnaire, participants were asked to provide demographic information.

Survey demographics. We recruited a total of 100 participants. Ages of participants were between 18 to 71 (μ=36.2, σ=11.4), and 41 of them are female (41%). Thirteen percent of our participants had a high school diploma (n=13); 36% had some college or associate degree (n=36); 36% had bachelor's degree (n=36); and 15% had a graduate or professional degree (n=15). The average time to take the survey was 11.7 minutes. The participants were remunerated for their participation in the survey.

4.2 Results

The analysis shows that the internal reliability of the 20-item smartphone-SeBIS is below the recommended cutoff point by Nunnally (1978) (Cronbach's α=.67<.70) [48]. We further conducted Confirmatory Factor Analysis (CFA) to examine whether our measurement of the construct is consistent with SeBIS by the goodness-of-fit of data to the latent variable model. We used several tests to determine the goodness-of-fit of data to the model of SeBIS, including the Comparative Fit Index (CFI) and the Root Mean Square Error of Approximation (RMSEA).

According to our results, the CFI and TLI were 0.565 and 0.490, which are below the cutoff (0.90) recommended by Netemeyer et al. [47]. Furthermore, our RMSEA and SRMR were 0.127 and 0.152 respectively, which are above the recommended cutoff points (a cutoff of 0.06 for RMSEA and 0.08 for SRMR [33])). These results indicate poor goodness-of-fit of our data to smartphone-SeBIS. Put another way: *the revised four dimensions of smartphone-SeBIS might not be the best fit for assessing users' smartphone security behavior intentions.*

5 Phase-2: Developing SSBS

5.1 Survey design and item generation

To develop a new scale to measure users' smartphone security behavior intentions we employed the approach used by Egelman and Peer (2015). We first generated a list of smartphone security behaviors and collected data on Amazon MTurk. We then conducted an Exploratory Factor Analysis (EFA) to extract the effective items for assessing users' smartphone security behavior.

Item generation. According to Egelman and Peer [20], the metric of security behavior should be "applicable" to and "widely accepted" by the majority of users. Therefore, we generated a list of different smartphone security behavior based on the views of security professionals.

In conducting our study, we invited a panel of 35 subject matter experts in the field of security. The panelists consisted of faculty members, graduate and undergraduate students specializing in cyber-security, who are participating in a security focused reading seminar. These expert panelists were tasked with identifying and ranking the 10 most critical types of smartphone security behavior(Table: 9). Meanwhile, two security researchers then categorized these security behaviors

into *Technical* and *Social* behaviors. The researchers also examined public security advice for smartphone security by the United States Computer Emergency Readiness Team (US-CERT) to ensure no important behaviors were missing from the list. Five security experts went through the list to determine if any item violated the principles of applicability and acceptance. Our initial list contained 45 types of behavior. We proceeded to translate these behaviors into personal statements. Survey participants were asked to read and rate each statement on a five point scale of frequency (From 'Never' to 'Always').

Survey Demographics. We collected 487 responses via Amazon MTurk. This is a larger sample than the sample size recommended by Hair *et al.* (minimum of 5 participants per item) [30]. The average age of participants was 34.6 and 44.8% were female (41%). About 11% of our participants had high school diplomas (n=54); 29% had some college or associate degree (n=142); 49.5% had a bachelor's degree (n=241); and 10.3% had a graduate or professional degree (n=50). The average time taken to complete the survey was 6.3 minutes. Participants were paid $0.75 after completing the survey and passing the validation checks.

5.2 Results

5.2.1 Exploratory Factor Analysis

Our analysis of Kaiser-Meyer-Olkin (KMO) test was 0.92 indicating the high sampling adequacy of variables, which suggest suitability for further factor analysis. Considering a large set of items, our approach was to refine our scales until the loading of each item was above 0.5 and was twice more than its loading on other components after a Varimax rotation [54]. Furthermore, we used optimal coordinates to determine the optimal number of factors, which is a non-graphical approach for factor determination [52]. The optimal coordinate is a determined point where the predicted eigenvalue is not greater than or equal to the mean eigenvalue by performing linear regression analysis of the last and $(i+1)$th eigenvalue [52]. By using optimal coordinates, we could overcome a limitation of subjective and unclear decision-making about the number of components to retain [52].

We performed three rounds of EFA to finalize our scale of smartphone security behavioral intention. In our first round of EFA, we first conducted Principal Component Analysis (PCA) and extracted five components in EFA based on optimal coordinate analysis. Next, we excluded 27 items based on the aforementioned loading criteria. In the second round of EFA, we followed the same procedure and extracted three components in EFA. 3 items were excluded from the list of items. In the third round of EFA, we also followed the same procedure performed in the last two rounds, extracted 2 components in EFA, and retained the remaining 14 items. The final set of items and their rotated factor loadings are presented in Table 3.

Analysis of the items revealed two distinct themes: *technical* approaches (*e.g.*, using a VPN and an anti-virus app) and *social* approaches (*e.g.*, verifying the source of texts before sharing and deleting suspicious communication). The *technical* items (T1...T8) describe actions that an Android user can take that are either supported by the underlying smartphone technology or can be supported by third-party add-on technology. For instance, the ability to reset the Advertising ID through the phone's Settings is an example of a feature supported by the underlying smartphone technology, while installing an anti-virus app is an example of a add-on feature by third-party. In contrast, the *social* items (S1...S8) pertain to behaviors that are socially constructed, in other words, it refers to behaviors that the user may exhibit while interacting and engaging with the technology. For example, item S2 refers to the user checking the source of an app during the process of downloading it. This is not a technical measure rather an interaction with the environment or context. Hence, our scale includes *Technical* and *Social* as two subscales.

5.2.2 Reliability of the Scale

We adopted the same approach used by Egelman and Peer [20] to examine the reliability of the scale based on three metrics. We first employed Cronbach's α, which is commonly used to assess internal consistency of a group of items. As shown in Table 3, the Cronbach's α for the full scale was 0.80. For subscales of technical and social approaches were 0.84 and 0.79 respectively. Our scale met the criteria of internal consistency that requires both full scale and all subscales to be above 0.7 [43, 49]. We subsequently leveraged the item-total correlation (ITC), which is the Pearson correlation between each item and the mean of all other items. All of our items' ITC are above the recommended threshold of 0.2 [21].

While assessing the reliability of the scale, it is also important to examine the diversity of the items of a scale and prevent the redundancy of the items [3]. Toward this end, we computed the average inter-item correlation (IIC) that not only evaluates the internal consistency but also tests the degree of redundancy of a set of items on a scale [10, 51]. The recommended correlational coefficient of IIC is between 0.20 and 0.40, which suggests that the items contain sufficient diversity of variance while they are still representative of the same construct [51]. The ITC of both our subscales fall within the range, which indicates the adequate level between consistency and diversity. Based on these three metrics, our full scale and sub-scales exhibit high reliability.

5.2.3 Convergent Validity: Correlation with SeBIS

To ensure that we assess the construct of users' security behavior, we measured the convergent validity of our scale and SeBIS. Convergent validity is a type of criterion validity that evaluates if a developed scale measures the same construct

Table 1: Pearson's Correlation between SeBIS and SSBS

	Correlation coefficient (p-value)	
SeBIS / SSBS	Technical approach	Social approach
Device securement	-.017 (p=.896)	.060 (p=.628)
Password generation	.290 (p=.018)	.229 (p=.064)
Proactive awareness	-.090 (p=.471)	.614 (p<.0001)
Update	.301 (p=.014)	.431 (p=.0003)

of the 'criterion' scale. We used SeBIS as our criterion because it is the only measure with high reliability for assessing security behavioral intentions. We collected a new dataset with 66 participants who completed both SeBIS and Smartphone Security Behavior Scale (SSBS). Then we conducted Pearson's correlation between SeBIS and SSBS. The average score of SeBIS had a significantly positive correlation with the average score of SSBS (r=.403, p=.0008). In addition, results show the positive significant correlation between the subscales of SeBIS and SSBS (see Table 1). These findings suggest that *participants who showed higher intentions in protecting their general security were also more likely to protect their smartphone security*. This confirms that our scale is measuring a similar construct with SeBIS, that of security behavior intentions.

5.2.4 Confirmatory data analysis

Our final step was to examine the goodness of fit of SSBS with the hypothesized latent components by performing Confirmatory Factor Analysis(CFA). We collected a new dataset with 358 U.S. participants from Amazon MTurk in the final round of our survey. Each participant was compensated for completing the survey. In order to mitigate the potential priming effect on participants' responses, we employed the same approach as in our phase-1 study by advertising the survey as research related to users' mobile phone usage and mental wellbeing so we included few related questions in the survey questionnaire. After completing the questionnaire, participants were asked to provide demographic information. To control for potential order effects on participants' responses, all survey sections were randomized. Additionally, we implemented a range of validity check measures to ensure data quality. These measures included the inclusion of several attention check questions throughout the survey to make sure the survey participants were attentive, restricting the survey participants to be between 18 and 65, including only 100% completed responses in our analysis, reverse coding when appropriate, and randomizing the order of survey questions to minimize any potential biases due to the ordering effects. In terms of demographics, 38% (n=136) of our participants were female and the average age of participants was 35.3 (σ=10.6). Each participant was paid $1 after completing the survey.

The reliability of full SSBS was 0.79, 0.81 for the *Technical* subscale, and 0.85 for the *Social* subscale. We conducted PCA with a Varimax rotation and extracted two components. The results show that all items were loaded on the same unique component as found in the previous EFA. We conducted CFA to examine the goodness-of-fit of the two-component model for users' smartphone security behavior intentions. We used the same approach employed in the Phase-1 study, by performing multiple test to determine the goodness of fit of our data to the model, including Comparative Fit Index (CFI), the Tucker-Lewis Index (TLI), the Root Mean Square Error of Approximation (RMSEA), and the Standardized Root Mean Square Residual (SRMR). Based on the analysis, the CFI and TLI were 0.954 and 0.942, which are above the cutoff (0.90) recommended by Netemeyer et al. [47]. Additionally, the RMSEA and SRMR were 0.054 and 0.059 respectively. Both scores are below the cutoff points recommended by [33]. Our results show a well goodness-of-fit of our data to our hypothesized two-component model. We also performed Pearson's correlation between the two subscales and found no significant correlations. Please see Table 2 for the details of CFA results.

6 Discussion & Future Work

6.1 Applications and Role of the SSBS

In this study, we determined that the psychological construct of smartphone security behavior differs from general security behavior measured by SeBIS [20]. Driven by this finding, we used a series of factor analyses to create a Smartphone Security Behavior Scale (SSBS) with 14 questions that loads onto two factors: a technical approach (using technical strategies to protect smartphones) and a social approach (being contextually aware and cautious while using their smartphones). Our scale exhibited satisfactory psychometric properties: the full scale and both the subscales have high internal consistency, all items map uniquely on one single component, and no correlation exists between the subscales while establishing convergent validity between SSBS and SeBIS. These indicate that SSBS is a well-established psychological construct and measurement. Furthermore, we distinguished a psychological construct of smartphone security behavior from a general security behavior measured by SeBIS [20]. This finding also corroborates that users have different security and privacy concerns and behaviors toward smartphone and laptop [9].

Using SSBS to measure Smartphone Security Behavior Intentions. Our new scale of smartphone security behavior intentions can be employed for various purposes. The most obvious utilization is for measuring smartphone end-users' security behavior intentions. While SeBIS has been shown to predict secure locking behavior on smartphones, some of its wording is outdated [22] as well some of its items are irrelevant to smartphone security actions. By contrast, SSBS items are specific to current smartphone functionality and

Table 2: SSBS Average variance extracted for CFA factor

	Standardized loading	R^2
Technical		
I reset my Advertising ID on my smartphone.	.715	.511
I hide device in my smartphone's bluetooth settings.	.641	.411
I change my passcode/PIN for my smartphone's screen lock at a regular basis.	.792	.627
I manually cover my smartphone's screen when using it in the public area (e.g., bus or subway).	.595	.354
I use an adblocker on my smartphone.	.529	.280
I use an anti-virus app.	.536	.287
I use a Virtual Private Network (VPN) app while connected to a public network.	.577	.333
I turn off WiFi on my smartphone when not actively using it.	.372	.139
Social		
I care about the source of the app when performing financial and/or shopping tasks on that app.	.770	.593
When downloading an app, I check that the app is from the official/expected source.	.785	.616
Before downloading a smartphone app I ensure the download is from official application stores (e.g. Apple App Store, GooglePlay, Amazon Appstore).	.799	.639
I verify the recipient/sender before sharing text messages or other information using smartphone apps.	.651	.423
I delete any online communications (i.e., texts, emails, social media posts) that look suspicious.	.651	.424
I pay attention to the pop-ups on my smartphone when connecting it to another device (e.g. laptop, desktop).	.578	.334

SSBS reduces the number of items in the scale. This makes SSBS valuable in environments at risk that exhibit high use of smartphone technology.

Our scale has numerous potential applications across a variety of contexts. For instance, in a healthcare setting, a doctor who wishes to utilize health-related apps for treatment or self-management may use our scale to determine whether her patients require educational interventions before using the app. In a workplace setting, employers can use our scale to evaluate the risk of accidental insider threats arising from employees' use of smartphones and implement interventions to promote more secure behavior. In an educational context, schools can deploy our scale to assess the smartphone security behavior of both teachers and students as they embrace the use of smartphones for online education. Schools may also use our scale to gauge the vulnerability of their students and faculty to potential cyberthreats through smartphones, such as cyberbullying and stalking. Additionally, enterprises can leverage our scale to design personalized and subject-based cybersecurity educational programs for training and onboarding their employees.

Moreover, researchers may utilize SSBS to investigate how behavior intentions change among different cultures and languages (similar to Sharif et al. [55] for SeBIS), or over time with educational or motivational interventions. For instance, researchers who are interested in smartphone malware prevention may use SSBS to explore the effect of smartphone security behavior intentions to vulnerability exploits.

Role of SSBS in modeling Smartphone Security Behavior. Davis et al. [24] in their *Theory of Reasoned Action*, postulated that behavior intention is an antecedent to behavior. SSBS can thus add to the predictive value of a computational model of smartphone security behavior targeting early interventions that seek to prevent security breaches stemming from smartphone attack entry-points.

Theory of Reasoned Action. The Theory of Reasoned Action also posits that security behavior intention is a function of behavioral (attitudes) and normative beliefs. These beliefs influence intentions through attitudes and/or subjective norms. The *Theory of Planned Behavior* further argues that beliefs are not purely volitional but related to acquired resources and opportunities for performing the given behavior. Studies and ensuing causal models on what and to what extent factors (beliefs) affect smartphone security behavior intentions and behavior can further enhance an SSBS-based framework. Lastly, SSBS can contribute together with SEBIS [20] and SA-6 [22] into a more general framework modeling behavior

Table 3: Factor loadings and reliability statistics of finalized scale

ID	Item	Technical	Social	Inter-total correlation
T1	I reset my Advertising ID on my smartphone.	.787		0.52
T2	I hide device in my smartphone's bluetooth settings.	.639		0.47
T3	I change my passcode/PIN for my smartphone's screen lock at a regular basis.	.629		0.51
T4	I manually cover my smartphone's screen when using it in the public area (e.g., bus or subway).	.621		0.55
T5	I use an adblocker on my smartphone.	.614		0.51
T6	I use an anti-virus app.	.612		0.53
T7	I use a Virtual Private Network (VPN) app while connected to a public network.	.604		0.42
T8	I turn off WiFi on my smartphone when not actively using it.	.544		0.47
S1	I care about the source of the app when performing financial and/or shopping tasks on that app.		.723	0.24
S2	When downloading an app, I check that the app is from the official/expected source.		.677	0.36
S3	Before downloading a smartphone app I ensure the download is from official application stores.		.677	0.21
S4	I verify the recipient/sender before sharing text messages or other information using smartphone apps.		.609	0.41
S5	I delete any online communications (i.e., texts, emails, social media posts) that look suspicious.		.552	0.25
S6	I pay attention to the pop-ups on my smartphone when connecting it to another device (e.g. laptop, desktop).		.526	0.39
2*	**Cronbach's alpha**	0.84	0.79	
	Inter-item correlation	0.40	0.39	

across different device types.

6.2 Limitations and Future Work

Scale Development Methodology. Our scale development process relies on security experts for generating the questionnaire items. There are other methods that can be used to incorporate expert opinions such as focus groups or Delphi studies [14]. Focus groups suffer from biases extending from individuals dominating the group opinion, which Delphi studies eliminate by collecting anonymized responses from experts through questionnaires, a process repeated for multiple rounds until consensus is reached. However, there is no element of discussion involved in Delphi, which runs the risk of vanishing opinion semantics. Additionally, the methodological process of a Delphi study is not well-established with numerous works illustrating Delphi variations with unclear reliability results. Instead, we adapted the same approach used by Egelman and Peer [20] and followed the 4-step scale development process proposed by Netemeyer et al. [47]. This approach has already been applied in the context of security behavior intentions to yield good reliability and predictive ability [19, 20].

Construct Validity. We performed established tests and demonstrated the reliability of SSBS and its goodness of fit with its two components. However, to better understand smartphone security behavior intentions, future work could further explore the convergence of SSBS with other related variables and its divergence from variables unrelated to security. We approached the problem primarily from a security standpoint, with privacy considerations only being secondary. More work is needed to understand socio-technical smartphone privacy behaviors.

Predicting Actual Behavior from Intentions. Lastly, intentions do not always result in behavior actions. The TRA and TPB theories come with limitations that SSBS inherits. For example, the TPB does not address the timeframe between a behavioral intention and an ensuing action and how this relationship can change over time. Moreover, the effects of other variables such as fear and threat of past experiences can further influence intentions. More work is needed to analyze how such factors influence the predictive power of SSBS. We plan to examine whether SSBS can predict relevant behavior actions and factors that affect that relationship in future work. Lastly, our findings indicate that a direct translation of SeBIS to the smartphone domain exhibits a poor goodness of fit. However, this should not be interpreted as SeBIS not being useful in measuring smartphone security intentions. In fact, Egelman, Harbach, and Peer found that SeBIS can predict smartphone secure screen locking behavior [19]. Our findings support the need for a new specialized measure if smartphone-specific wording is preferred or necessitated by the application context.

Comprehensiveness of Scale. While we followed an established psychometric process to operationalize smartphone security intentions, the comprehensiveness of the resulting scale should be further evaluated. This can be conducted through questionnaires with smartphone security experts to reveal whether the items can comprehensively cover the entire or a large portion of the spectrum of security behavior intentions. Such a study could reveal important items that have not been considered in our study.

Our paper also only focused on understanding the security facets of the smartphone; this work can be used as motivation for understanding privacy behavior. Lastly, studies with users could further establish user comprehension of the items' wording.

Demographic Characterization. Like the majority of psychometric measurements, SSBS is based on self-reports. To reduce potential social desirability bias, we took several precautionary procedures: being careful about wording questions in a non-judgmental way, making the surveys anonymous, and keeping the purpose of each survey vague. For data cleaning, we excluded unattended responses [53]. Moreover, our results

might not generalize since our sample is based entirely on US-based Amazon Mechanical Turk workers. Further studies are needed to support our findings across different cultures, languages, and social norms.

Average Variance Extracted(AVE). R^2 represents the proportion of variance in each item that is explained by the factor, and a desirable value for R^2 is at least 0.30. As shown in Table 2, the last technical item (T8) had a low R^2 value of 0.139, indicating that it did not capture the behavior well. This could be due to the users' uncertainty about how disconnecting from a WiFi network could enhance security, rather than using a VPN when connected to an insecure network. To calculate AVE, which is the average of R^2 values of items, a value of 0.50 or higher is recommended. If excluding T8 from the calculation, the AVE for Technical items was 0.401, which was slightly below the suggested threshold, while the AVE for Social items was 0.505, which met the criterion. The Technical scale has its merit and potential, as it is based on a rigorous literature review and empirical data collection, and it reflects some aspects of users' technical self-efficacy that are relevant for smartphone usable security behaviors. However, we also acknowledge the limitation of the low Technical AVE value and propose this as a challenge for future research in developing and validating a technical smartphone scale. We encourage future researchers to use our results as a reference point for addressing this limitation.

7 Conclusion

In this study, we found that smartphone security behavior differs from general security behavior. We thus carried out a series of factor analyses to create a Smartphone Security Behavior (Intentions) Scale (SSBS) with 14 questions that load onto two factors: technical (using technical strategies to protect smartphones) and social (being contextually cautious while using smartphones). Our scale exhibited satisfactory psychometric properties: the full scale and both the subscales have high internal consistency, all items map uniquely on one single component, and no correlation exists between the subscales. We established convergent validity between SSBS and an existent well-established security behavior measurement, SeBIS. These results support demonstrate SSBS can be a valuable specialized instrument in our arsenal for better understanding human smartphone security behavior, especially for security researchers and HCI designers that hope to preserve such cybersecurity becomes even more prevalent. Nevertheless, we recognize that the technical factor of the scale has a moderate fit and could be improved by further refinement. This is a limitation of our study that we plan to address in future work.

Acknowledgments

This work was supported in part by NSF CNS 19-55228 (SPLICE). The views expressed are those of the authors only.

References

[1] Icek Ajzen et al. The theory of planned behavior. *Organizational behavior and human decision processes*, 50(2):179–211, 1991.

[2] Bakheet Aljedaani, Aakash Ahmad, Mansooreh Zahedi, and M Ali Babar. Security awareness of end-users of mobile health applications: an empirical study. In *MobiQuitous 2020-17th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*, pages 125–136, 2020.

[3] Mary J Allen and Wendy M Yen. *Introduction to measurement theory*. Waveland Press, 2001.

[4] Hazim Almuhimedi, Florian Schaub, Norman Sadeh, Idris Adjerid, Alessandro Acquisti, Joshua Gluck, Lorrie Faith Cranor, and Yuvraj Agarwal. Your location has been shared 5,398 times!: A field study on mobile app privacy nudging. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pages 787–796. ACM, 2015.

[5] Antonio Bianchi, Jacopo Corbetta, Luca Invernizzi, Yanick Fratantonio, Christopher Kruegel, and Giovanni Vigna. What the app is that? deception and countermeasures in the android user interface. In *2015 IEEE Symposium on Security and Privacy*, pages 931–948. IEEE, 2015.

[6] Andrew bunnie Huang. Betrusted: Improving security through physical partitioning. *IEEE Pervasive Computing*, 19(02):13–20, 2020.

[7] Bobby J Calder, Lynn W Phillips, and Alice M Tybout. The concept of external validity. *Journal of consumer research*, 9(3):240–244, 1982.

[8] Pew Research Center. Mobile fact sheet. https://www.pewinternet.org/fact-sheet/mobile/, June 2022.

[9] Erika Chin, Adrienne Porter Felt, Vyas Sekar, and David Wagner. Measuring user confidence in smartphone security and privacy. In *Proceedings of the eighth symposium on usable privacy and security*, page 1. ACM, 2012.

[10] Ronald Jay Cohen and Mark E Swerdlik. *Psychological testing and assessment 6E*. New York: McGraw Hill, 2005.

[11] Ronald Jay Cohen, Mark E Swerdlik, and Suzanne M Phillips. *Psychological testing and assessment: An introduction to tests and measurement*. Mayfield Publishing Co, 1996.

[12] Jose M Cortina. What is coefficient alpha? an examination of theory and applications. *Journal of applied psychology*, 78(1):98, 1993.

[13] Lee J Cronbach. Coefficient alpha and the internal structure of tests. *psychometrika*, 16(3):297–334, 1951.

[14] Norman Dalkey and Olaf Helmer. An experimental application of the delphi method to the use of experts. *Management science*, 9(3):458–467, 1963.

[15] Amit Das and Habib Ullah Khan. Security behaviors of smartphone users. *Information & Computer Security*, 24(1):116–134, 2016.

[16] Jamie Davies. *Infographic: What do we actually use our smartphones for?*, July 2017. http://telecoms.com/483334/infographic-what-do-we-actually-use-our-smartphones-for.

[17] Fred D Davis. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS quarterly*, pages 319–340, 1989.

[18] Fred D Davis, Richard P Bagozzi, and Paul R Warshaw. User acceptance of computer technology: A comparison of two theoretical models. *Management science*, 35(8):982–1003, 1989.

[19] Serge Egelman, Marian Harbach, and Eyal Peer. Behavior ever follows intention?: A validation of the security behavior intentions scale (sebis). In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 5257–5261. ACM, 2016.

[20] Serge Egelman and Eyal Peer. Scaling the security wall: Developing a security behavior intentions scale (sebis). In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 2873–2882. ACM, 2015.

[21] Brian S Everitt and Anders Skrondal. *The Cambridge dictionary of statistics*. New York University, 2010.

[22] Cori Faklaris, Laura A Dabbish, and Jason I Hong. A self-report measure of end-user security attitudes (sa-6). In *Fifteenth Symposium on Usable Privacy and Security ({SOUPS} 2019)*, 2019.

[23] Adrienne Porter Felt, Elizabeth Ha, Serge Egelman, Ariel Haney, Erika Chin, and David Wagner. Android permissions: User attention, comprehension, and behavior. In *Proceedings of the eighth symposium on usable privacy and security*, page 3. ACM, 2012.

[24] Martin Fishbein. A theory of reasoned action: some applications and implications. 1979.

[25] Drew Fisher, Leah Dorner, and David Wagner. Short paper: location privacy: user behavior in the field. In *Proceedings of the second ACM workshop on Security and privacy in smartphones and mobile devices*, pages 51–56. ACM, 2012.

[26] Yanick Fratantonio, Chenxiong Qian, Simon P Chung, and Wenke Lee. Cloak and dagger: from two permissions to complete control of the ui feedback loop. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 1041–1057. IEEE, 2017.

[27] Alisa Frik, Juliann Kim, Joshua Rafael Sanchez, and Joanne Ma. Users' expectations about and use of smartphone privacy and security settings. In *CHI Conference on Human Factors in Computing Systems*, pages 1–24, 2022.

[28] Hugo Gascon, Sebastian Uellenbeck, Christopher Wolf, and Konrad Rieck. Continuous authentication on mobile devices by analysis of typing motion behavior. *Sicherheit 2014–Sicherheit, Schutz und Zuverlässigkeit*, 2014.

[29] Ceenu George, Daniel Buschek, Andrea Ngao, and Mohamed Khamis. Gazeroomlock: Using gaze and head-pose to improve the usability and observation resistance of 3d passwords in virtual reality. In *International Conference on Augmented Reality, Virtual Reality and Computer Graphics*, pages 61–81. Springer, 2020.

[30] Joseph F Hair, William C Black, Barry J Babin, and Rolph E Anderson. *Multivariate data analysis: Pearson new international edition*. Pearson Higher Ed, 2013.

[31] Marian Harbach, Alexander De Luca, and Serge Egelman. The anatomy of smartphone unlocking: A field study of android lock screens. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 4806–4817. ACM, 2016.

[32] Marian Harbach, Emanuel Von Zezschwitz, Andreas Fichtner, Alexander De Luca, and Matthew Smith. It'sa hard lock life: A field study of smartphone (un) locking behavior and risk perception. In *10th Symposium On Usable Privacy and Security ({SOUPS} 2014)*, pages 213–230, 2014.

[33] Li-tze Hu and Peter M Bentler. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling: a multidisciplinary journal*, 6(1):1–55, 1999.

[34] Beth H Jones and Amita Goyal Chin. On the efficacy of smartphone security: a critical analysis of modifications in business students' practices over time. *International*

Journal of Information Management, 35(5):561–571, 2015.

[35] Joon-Myung Kang, Sin-seok Seo, and James Won-Ki Hong. Usage pattern analysis of smartphones. In *2011 13th Asia-Pacific Network Operations and Management Symposium*, pages 1–8. IEEE, 2011.

[36] Patrick Gage Kelley, Lorrie Faith Cranor, and Norman Sadeh. Privacy as part of the app decision-making process. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 3393–3402. ACM, 2013.

[37] Murat Koyuncu and Tolga Pusatli. Security awareness level of smartphone users: An exploratory case study. *Mobile Information Systems*, 2019, 2019.

[38] Shakuntala P Kulkarni and Sachin Bojewar. Vulnerabilities of smart phones. *International Research Journal of Engineering and Technology*, 2(9):2422–2426, 2015.

[39] Mengyuan Li, Yan Meng, Junyi Liu, Haojin Zhu, Xiaohui Liang, Yao Liu, and Na Ruan. When csi meets public wifi: Inferring your mobile phone password via wifi signals. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 1068–1079. ACM, 2016.

[40] Rui Li, Wenrui Diao, Zhou Li, Jianqi Du, and Shanqing Guo. Android custom permissions demystified: From privilege escalation to design shortcomings. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 70–86. IEEE, 2021.

[41] Huigang Liang and Yajiong Xue. Understanding security behaviors in personal computer usage: A threat avoidance perspective. *Journal of the association for information systems*, 11(7):394–413, 2010.

[42] Thomas J Madden, Pamela Scholder Ellen, and Icek Ajzen. A comparison of the theory of planned behavior and the theory of reasoned action. *Personality and social psychology Bulletin*, 18(1):3–9, 1992.

[43] Robert K McKinley, Terjinder Manku-Scott, Adrian M Hastings, David P French, and Richard Baker. Reliability and validity of a new measure of patient satisfaction with out of hours primary medical care in the united kingdom: development of a patient questionnaire. *Bmj*, 314(7075):193, 1997.

[44] Joel Michell. Is psychometrics pathological science? *Measurement*, 6(1-2):7–24, 2008.

[45] Kireet Muppavaram, Meda Sreenivasa Rao, Kaavya Rekanar, and R Sarath Babu. How safe is your mobile app? mobile app attacks and defense. In *Proceedings of the Second International Conference on Computational Intelligence and Informatics*, pages 199–207. Springer, 2018.

[46] Alexios Mylonas, Anastasia Kastania, and Dimitris Gritzalis. Delegate the smartphone user? security awareness in smartphone platforms. *Computers & Security*, 34:47–66, 2013.

[47] Richard G Netemeyer, William O Bearden, and Subhash Sharma. *Scaling procedures: Issues and applications*. Sage Publications, 2003.

[48] Jum C. Nunnally. *Psychometric theory / Jum C. Nunnally*. McGraw-Hill New York, 2d ed. edition, 1978.

[49] Jum C Nunnally and Ira Bernstein. *Psychometric theory 3E*. Tata McGraw-Hill Education, 1994.

[50] oberlo.com. How many people have smartphones in 2020?, 2020. https://www.oberlo.com/statistics/how-many-people-have-smartphones.

[51] Ralph L Piedmont. Inter-item correlations. *Encyclopedia of quality of life and well-being research*, pages 3303–3304, 2014.

[52] Gilles Raîche, Theodore A Walls, David Magis, Martin Riopel, and Jean-Guy Blais. Non-graphical solutions for cattell's scree test. *Methodology*, 2013.

[53] Elissa M Redmiles, Sean Kross, Alisha Pradhan, and Michelle L Mazurek. How well do my results generalize? comparing security and privacy survey results from mturk and web panels to the us. Technical report, 2017.

[54] Gerard Saucier. Mini-markers: A brief version of goldberg's unipolar big-five markers. *Journal of personality assessment*, 63(3):506–516, 1994.

[55] Yukiko Sawaya, Mahmood Sharif, Nicolas Christin, Ayumu Kubota, Akihiro Nakarai, and Akira Yamada. Self-confidence trumps knowledge: A cross-cultural study of security behavior. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 2202–2214. ACM, 2017.

[56] Bingyu Shen, Lili Wei, Chengcheng Xiang, Yudong Wu, Mingyao Shen, Yuanyuan Zhou, and Xinxin Jin. Can systems explain permissions better? understanding users' misperceptions under smartphone runtime permission model. In *USENIX Security Symposium*, pages 751–768, 2021.

[57] Sooel Son, Daehyeok Kim, and Vitaly Shmatikov. What mobile ads know about mobile users. In *NDSS*, 2016.

[58] Mohsen Tavakol and Reg Dennick. Making sense of cronbach's alpha. *International journal of medical education*, 2:53, 2011.

[59] Nik Thompson, Tanya Jane McGill, and Xuequn Wang. "security begins at home": Determinants of home computer and mobile device security behavior. *computers & security*, 70:376–391, 2017.

[60] Güliz Seray Tuncay, Soteris Demetriou, Karan Ganju, and Carl A Gunter. Resolving the predicament of android custom permissions. In *Proceedings of Network and Distributed System Security (NDSS) Symposium*, 2018.

[61] Güliz Seray Tuncay, Jingyu Qian, and Carl A Gunter. See no evil: phishing for permissions with false transparency. In *Proceedings of the 29th USENIX Conference on Security Symposium*, pages 415–432, 2020.

[62] Silas Formunyuy Verkijika. Understanding smartphone security behaviors: An extension of the protection motivation theory with anticipated regret. *Computers & Security*, 77:860–870, 2018.

[63] Jice Wang, Yue Xiao, Xueqiang Wang, Yuhong Nan, Luyi Xing, Xiaojing Liao, JinWei Dong, Nicolas Serrano, Haoran Lu, XiaoFeng Wang, et al. Understanding malicious cross-library data harvesting on android. In *USENIX Security Symposium*, pages 4133–4150, 2021.

[64] Primal Wijesekera, Arjun Baokar, Ashkan Hosseini, Serge Egelman, David Wagner, and Konstantin Beznosov. Android permissions remystified: A field study on contextual integrity. In *24th {USENIX} Security Symposium ({USENIX} Security 15)*, pages 499–514, 2015.

[65] Luyi Xing, Xiaorui Pan, Rui Wang, Kan Yuan, and XiaoFeng Wang. Upgrading your android, elevating my malware: Privilege escalation through mobile os updating. In *2014 IEEE symposium on security and privacy*, pages 393–408. IEEE, 2014.

[66] Wu Zhou, Yajin Zhou, Xuxian Jiang, and Peng Ning. Detecting repackaged smartphone applications in third-party android marketplaces. In *Proceedings of the second ACM conference on Data and Application Security and Privacy*, pages 317–326. ACM, 2012.

Appendices

8 Translating SeBIS to the smartphone domain

Smartphone-SeBIS is based on the four dimensions of SeBIS: device securement, password management, proactive awareness, and update (Table 4). We generated items by revising SeBIS's through *word/phrase substitution*, *word/phrase revision*, *item deletion*, *item addition*. The resulting scale is depicted in Table 5.

9 Common Method Bias Test

To test if the Common Method Bias (CMB) (7)existed in the mode, we adopted the Harman Single Factor approach. We conducted exploratory factor analysis where all variables are loaded onto one factor. According to the result, the Harman Single Factor technique estimates the common method variance to be 26.85% which is below the commonly accepted threshold of 50%; this suggests that common method bias might not be a problem in the study.

Table 4: Items for the **original** Security Behavior Intentions Scale (SeBIS) and associated sub-scales.

Dimension	**Item**
Device Securement	I set my computer screen to automatically lock if I don't use it for a prolonged period of time.
	I use a password/passcode to unlock my laptop or tablet.
	I manually lock my computer screen when I step away from it.
	I use a PIN or passcode to unlock my mobile phone.
Password Generation	I do not change my passwords, unless I have to.
	I use different passwords for different accounts that I have.
	When I create a new online account, I try to use a password that goes beyond the site's minimum requirements.
	I do not include special characters in my password if it's not required.
Proactive Awareness	When someone sends me a link, I open it without first verifying where it goes.
	I know what website I'm visiting based on its look and feel, rather than by looking at the URL bar.
	I submit information to websites without first verifying that it will be sent securely (e.g., SSL, "https://", a lock icon).
	When browsing websites, I mouseover links to see where they go, before clicking them.
	If I discover a security problem, I continue what I was doing because I assume someone else will fix it.
Updating	When I'm prompted about a software update, I install it right away.
	I try to make sure that the programs I use are up-to-date.
	I verify that my anti-virus software has been regularly updating itself.

Table 5: Preliminary set of survey items developed based on SeBIS (*smartphone-SeBIS*)

Dim..	ID	Item	μ	σ
Device Securement	DS1	I use biometrics (fingerprint, face recognition) to unlock my smartphone.	2.41	1.6
	DS2	I enable encrypted storage on my smartphone.	2.41	1.44
	DS3	I use a rooted/jailbroken phone (r).	1.45	1.07
	DS4	I turn on the "lost my device" feature on my smartphone.	2.5	1.6
	DS5	I use a password/passcode to unlock my smartphone.	3.76	1.51
Password management	PM1	I regularly change my password for online services/accounts using my smartphone.	2.36	1.13
	PM2	I share my smartphone's passcode/PIN with other(s). (r)	1.51	0.99
	PM3	I use password manager app to manage my passwords on my smartphone.	1.88	1.31
Proactive awareness	PA1	When downloading an app, I check that the app is from the official/expected source.	3.95	0.99
	PA2	Before downloading a smartphone app I ensure the download is from official application stores (e.g. Apple App Store, GooglePlay, Amazon Appstore)	4.13	1.14
	PA3	I reset my Advertising ID on my smartphone.	1.6	1.02
	PA4	I manually revoke permissions from apps.	3	1.09
	PA5	I grant smartphone apps the permissions they request. (r)	3.2	0.85
	PA6	I disable geotagging of images captured by smartphone's camera app.	3.23	1.48
	PA7	I check which apps are running in the background.	3.33	1.14
	PA8	I check my smartphone's privacy settings.	3.31	1.17
	PA9	When receiving a link from an unknown source via SMS, I click the link immediately. (r)	1.67	1.04
Update	UP1	When I'm prompted about a software update on my smartphone, I install it right away.	3.3	1.13
	UP2	I make sure that the smartphone applications I use are up-to-date.	3.61	0.92

Table 6: Developed smartphone security behavior measurement

Research	Smartphone Security Behavior Measurement	Scale
Das and Khan [2016]	1. I lock my smartphone with a PIN or password. 2. I update my software when new versions are released. 3. I have installed a mobile anti-virus program. 4. I encrypt confidential information (e.g., passwords, bank details, ...) on my smartphone. 5. I avoid storing confidential information (e.g., passwords, bank details, ...) on my smartphone. 6. I review security features of apps before installing them on my smartphone.	6-point scale
Jones and Chin [2015]	1. Have you set the idle timeout (so that the screen goes dark) to a shorter time than the factor default? 2. To wake up after idle, is a password or other code required on your smartphone? 3. Do you disable Bluetooth when it's not in use? 4. Do you disable GPS (navigation) when you are not using it? 5. When you use your phone to connect to Wi-Fi wireless networks, do you only connect to encrypted password-protected networks? 6. Select one answer regarding anti0virus software: "Anti-virus software has been downloaded and installed on my phone and I use it..." 7. Select one answer regarding encryption software: "Encryption software has been downloaded and installed on my phone and I use it..."	5-point categorial scale (Frequently, Sometimes, Rarely/Never, Software not installed, Don't know)
Thompson et al. [2017]	1. I have installed security software on my device 2. I have recent backups of my device 3. I have enabled automatic updating of my computer software 4. I use security software (anti-virus/anti malware) 5. My device is secured by a password.	7-point Likert scale (Strongly Disagree-Strongly Agree)
Verkijika [2018]	1. I have installed security software on my device 2. I have recent backups of my device 3. I have enabled automatic updating of my computer software 4. I regularly use security software (anti-virus/anti malware) on my smartphone. 5. My smartphone is secured by a password or another authentication method (e.g., fingerprint).	5-point Likert scale (Strongly Disagree-Strongly Agree)

Table 7: Common Method Bias Test

Dimensions	Eigenvalue	Proportion (%)	Cumulative (%)
1	3.759	26.851	26.851
2	3.348	23.916	50.767
3	1.046	7.471	58.238
4	0.812	5.805	64.044
5	0.716	5.114	69.158
6	0.628	4.489	73.648
7	0.592	4.235	77.883
8	0.568	4.061	81.945
9	0.532	3.801	85.745
10	0.483	3.451	89.197
11	0.443	3.169	92.367
12	0.386	2.759	95.127
13	0.358	2.557	97.684
14	0.324	2.315	100.000

Table 8: Correlation between SSBS Technical (**T**) and Social (**S**) Scales

	T1	**T2**	**T3**	**T4**	**T5**	**T6**	**T7**	**T8**	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**
T1	1.000	.455	.594	.401	.327	.339	.460	.270	-.196	-.108	-.159	-.047	-.065	-.002
T2		1.000	.486	.405	.342	.286	.401	.344	.009	.024	.007	.118	.083	.112
T3			1.000	.484	.393	.432	.445	.282	-.017	.018	-.011	.106	.036	.102
T4				1.000	.302	.363	.282	.239	.004	.078	.062	.100	.141	.103
T5					1.000	.472	.352	.098	.045	.029	.001	.119	.066	.002
T6						1.000	.243	.169	.017	.043	.020	.120	.141	.029
T7							1.000	.199	-.033	.087	.033	.051	.037	.072
T8								1.000	.009	.053	.053	.083	.041	.122
S1									1.000	.612	.616	.473	.537	.420
S2										1.000	.629	.498	.515	.444
S3											1.000	.545	.484	.475
S4												1.000	.423	.401
S5													1.000	.381
S6														1.000

Table 9: List of Original Items of smartphone security behaviors generated by security professionals

ID	Item	μ	σ
A1	I turn off WiFi on my smartphone when not actively using it.	2.88	1.38
A2	I perform banking transactions/operations on my smartphone while connected to a public network.	3.62	1.35
A3	I connect to public WiFi using my smartphone.	2.92	1.78
A4	I use a Virtual Private Network (VPN) app while connected to a public network.	2.43	1.38
A5	When downloading an app, I check that the app is from the official/expected source.	3.92	1.07
A6	Before downloading a smartphone app I ensure the download is from official application stores (e.g. Apple App Store, GooglePlay, Amazon Appstore)	4.04	1.08
A7	I manually revoke permissions from apps.	3.15	1.11
A8	I grant smartphone apps the permissions they request.	2.56	0.89
A9	I disable geotagging of images captured by smartphone's camera app.	3.17	1.39
A10	I check which apps are running in the background.	3.59	1
A11	I delete apps I don't frequently use.	3.88	0.97
A12	I enable two-step authentication when offered by an app.	3.45	1.15
A13	I reset my Advertising ID on my smartphone.	2.32	1.36
A14	I check my smartphone's privacy settings.	3.5	1.06
A15	I turn off location services on my smartphone when I am not actively using them	3.41	1.28
A16	I turn off bluetooth (NFC, wifi) on my smartphone when I am not actively using it	3.67	1.32
A17	I use an anti-virus app	2.72	1.53
A18	I store proprietary business information on my smartphone.	3.7	1.33
A19	I store personal health information on my smartphone.	3.52	1.39
A20	I verify the recipient/sender before sharing text messages or other information using smartphone apps	3.71	1.14
A21	I use an adblocker on my smartphone.	2.84	1.48
A22	I pay attention to the pop-ups on my smartphone when connecting it to another device (e.g. laptop, desktop).	3.83	1.07
A23	I care about the source of the app when performing financial and/or shopping tasks on that app	4.05	0.98
A24	I back-up my smartphone's contacts, photos and videos on another device/cloud	3.44	1.23
A25	I delete any online communications (i.e., texts, emails, social media posts) that look suspicious	3.92	1.11
A26	I get permissions from my friends before sharing them on a photo or video online	3.48	1.24
A27	I check the latest news updates regarding my smartphone and apps	3.36	1.09
A28	I use private browsing on my smartphone.	3.06	1.16
A29	I use biometrics (fingerprint, face recognition) to unlock my smartphone.	3.07	1.61
A30	I enable encrypted storage (or phone memory) on my smartphone.	2.87	1.48
A31	I use a rooted/jailbroken phone.	1.97	1.35
A32	I turn on the "lost my device" feature on my smartphone.	2.89	1.53
A33	I use a password/passcode to unlock my smartphone.	3.87	1.27
A34	I hide device in my smartphone's bluetooth settings.	2.63	1.43
A35	I use a privacy screen on my smartphone	2.58	1.51
A36	I manually cover my smartphone's screen when using it in the public area (e.g., bus or subway).	2.89	1.25
A37	I change my password for online services/accounts using my smartphone.	2.89	1.24
A38	I share my smartphone's passcode/PIN with other(s).	4	1.27
A39	I use password manager app to manage my passwords on my smartphone.	2.55	1.48
A40	I store passwords and usernames on my smartphone.	3.32	1.41
A41	I change my passcode/PIN for my smartphone's screen lock at a regular basis.	2.7	1.31
A42	I use different passwords for different accounts that I have on my smartphone.	3.68	1.14
A43	When I'm prompted about a software update on my smartphone, I install it as soon as I can	3.53	1.09
A44	I make sure that the programs smartphone applications I use are up-to-date.	3.79	1.05
A45	Before downloading a smartphone app I read its privacy policy to ensure my information is handled securely	2.96	1.33

Privacy Mental Models of Electronic Health Records: A German Case Study

Rebecca Panskus[1], Max Ninow[2], Sascha Fahl[3], Karola Marky[1,2]
[1]*Ruhr-University Bochum, Germany,* [2]*Leibniz University Hannover, Germany*
[3]*CISPA Helmholtz Center for Information Security, Germany*

Abstract

Central digitization of health records bears the potential for better patient care, e.g., by having more accurate diagnoses or placing less burden on patients to inform doctors about their medical history. On the flip side, having electronic health records (EHRs) has privacy implications. Hence, the data management infrastructure needs to be designed and used with care. Otherwise, patients might reject the digitization of their records, or the data might be misused. Germany, in particular, is currently introducing centralized EHRs nationwide. We took this effort as a case study and captured the privacy mental models of EHRs. We present and discuss the findings of an interview study where we investigated expectations towards EHRs and perceptions of the German infrastructure. Most participants were positive but skeptical, yet expressed a variety of misconceptions, especially regarding data exchange with health insurance providers and read-write access to their EHRs. Based on our results, we make recommendations for digital infrastructure providers, such as developers, system designers, and healthcare providers.

1 Introduction

Centralized electronic health records (EHRs) bear the potential for providing better patient care [16], for instance, by having more accurate diagnoses, improved patient safety [32,42], and cost reduction [18]. Some countries deploy such centralized infrastructures, such as the UK [36], Denmark, or Australia.Yet, despite the apparent benefits, most countries do not have a digital infrastructure. Hence, sensible and important health-related data must be shared between healthcare practitioners, mostly by the patients. Introducing central, nationwide digitization of health-related data requires care. Otherwise, (a) patients might not adopt using the digital infrastructure [31,34] or (b) the digital infrastructure might be misused by malicious actors [6].

Research has repeatedly shown that privacy perceptions of patients play an integral role in the context of EHRs [22, 30,34]. Specific concerns included unauthorized access [30], misuse of data [31], or increased health insurance costs for patients with certain health conditions [5,22,22,30,31].

In this paper, we investigate the specific use case of Germany, where the Federal Ministry of Health introduced national EHRs in January 2021 [21]. Germany uses an infrastructure that turns health insurance companies into providers of EHR access apps. However, they can only access specific data, which introduces challenging trust assumptions toward insurance companies. So far, using EHRs is voluntary for patients, and most German citizens are not even aware of EHRs, yet many informed individuals would like to use it [1]. Further, Germany's health infrastructure has a unique feature: Access to the centrally stored EHRs is controlled through mobile apps provided by health insurance companies [21] resulting in 85 different apps [23].

To contribute a part in evaluating the understanding and acceptance of German citizens towards EHRs, we first took a look at the mental models of individuals, which are internal representations that humans derive from the real world, e.g., how and why a technology works [25]. In related domains, such as the IoT, mental models profoundly impact adopting systems that handle sensitive health data [5,34]. This motivates our first research question:

RQ1: Which mental models do patients have regarding EHRs? – We specifically focus on privacy, data access, and trust including expectations about data handling.

Since the correctness of mental models also impacts adoption and usage intention [25], we specifically investigate misconceptions:

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.
USENIX Symposium on Usable Privacy and Security (SOUPS) 2023.
August 6–8, 2023, Anaheim, CA, USA

RQ2: What are patients' misconceptions of the EHR? – We specifically focus on the German EHR infrastructure.

Finally, we investigate risk perceptions of individuals:

RQ3: Which risks do patients perceive in the EHR context? – We specifically focus on the German EHR infrastructure.

To answer the above research questions, we conducted semi-structured interviews with 21 participants that included drawing mental models of the expected German infrastructure. We also confronted participants with the actual national infrastructure to capture their perceptions.

Our results show that the participants consider health insurance companies to play a central role. However, we identified many misconceptions about that role that mostly originate from misconceptions already in the analog world. For instance, participants mistakenly thought that health insurance companies had detailed access to all patient data. Patients critically viewed the fact that health insurance companies, on the one hand, provide the apps to control EHRs, yet are, on the other hand, not allowed or should not be allowed to access all patient data. Most participants also considered it to have a rather negative impact that patients in Germany are allowed to add and delete documents. We demonstrate further expectations and misconceptions of the patients focusing on trust and privacy. Based on our findings, we leverage the lessons learned from this use case to inform infrastructure design, data handling, and access to EHR infrastructures.

Research contributions: In the course of this paper we make the following contributions:

1. **First mental model investigation of German EHR:** We present the first investigation of user perceptions of the German EHR. We specifically investigate the privacy mental models of 21 participants through semi-structured interviews and a drawing exercise.
2. **Analysis of perceived risks, expectations & misconceptions:** Among our results, we show perceived risks, misconceptions, and expectations towards EHR highlighting challenges arising from health insurance companies' unique role in the German infrastructure.
3. **Overall recommendations for EHRs:** We conclude with recommendations for digital infrastructure providers, such as developers, system designers, and healthcare providers. We further provide viable lessons learned from the German use case.

2 Background & Related Work

Electronic health records (EHRs) have several advantages compared to their analog counterpart, especially in terms of availability, completeness, and accessibility for different authorized stakeholders, the digital version presents a better solution for patients. There are two main ways to digitize patient records: (1) EHRs, as detailed above, and (2) personal EHRs managed by the patients (PEHRs), e.g., having the data on a USB drive. In this section, we first introduce the German infrastructure to store and access EHRs as defined by the Federal Ministry of Health [21]. Then, we detail related work on mental models, privacy, and investigations of (P)EHRs.

2.1 German Infrastructure

The German EHR was introduced in 2022 by the Patient Data Protection Act, which obliges all health insurance companies to provide all insured persons with an EHR upon request from the beginning of 2021.

Health Insurances in Germany: To understand the German infrastructure, we first need to provide information about German health care. Germany has two types of health insurance: (1) statutory health insurance companies and (2) private ones. Patients insured by statutory health insurance companies pay contributions calculated based on their income. Compared to that, the contributions for private health insurance depend on the age and the health of patients when they enter their contract. For statutory insured patients, their health insurance company pays most costs directly to the health care providers. Privately insured patients pay the health care providers themselves and get reimbursed from the insurance. In both cases, health insurances get a) the doctor ID, b) received treatment (incl. diagnosis & billing codes), and c) data to identify the patient [39], which is legally defined in German law. Consequently, health insurances have to follow binding rates that are specified nationwide. From a privacy perspective, the health insurance company cannot get more detailed information, and the received information can be used for billing purposes.

Data: The data is stored on a central server managed by *Gematik*, the National Agency for Digital Medicine. Patients can store their emergency data, medication plans, doctor's letters, findings, or X-rays in their EHR. They can also upload their data, e.g., their blood glucose diary. All data stored is protected by encryption. Patients can also delete data at any time. The data is lost if doctors do not have the data stored locally in their doctor's office.

Availability / Access: The EHR is available for patients that have an electronic health card. Those cards are distributed by statutory health insurance. Mobile device apps can manage the data access on a (1) smartphone, or (2) tablet, and even without possessing such a device, the EHR can be accessed via (3) the patient's electronic health card and a PIN. Using one of these three access methods ensures the availability of all existing documents when visiting a healthcare facility the patient has not been to before. Patients with private health

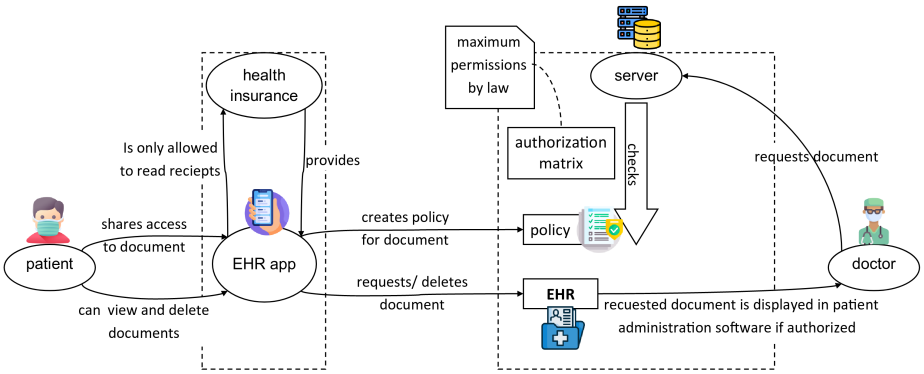


Figure 1: Schematic overview of the German infrastructure.

insurance, without an electronic health card, cannot use the EHR.The patients themselves determine who is granted access to the EHR. The patient defines who may access their data and to what extent it is shown to a specific entity. Further, patients can decide which data is uploaded into the EHR and delete data they do not want anymore.

App Provider: The respective health insurance of the patient functions as the provider of the access app. Thus, there is no central app for all German citizens, but as many apps as there are health insurance companies. However, the health insurance only provides the platform/infrastructure for managing the EHR and does not have access to the data. Some health insurances also provide desktop apps [23]. At the beginning of 2023, there were 85 different apps provided by health insurances [23].

2.2 Mental Models & Privacy

This section first introduces mental models and different investigations in the privacy context.

Mental Models & Investigations. Mental models are internal representations in the human mind used to explain the real world to decide on how to act [25]. Specifically, in the scope of mental models of technology, two model types are distinguished: (1) functional and (2) structural models [37]. The former (functional models) mean that individuals know how to use technology and its implications, yet detailed knowledge on *how* the technology works is not present. The latter (structural models) means that individuals have a detailed understanding of how technology works. Consequently, humans have more or less accurate and detailed mental models [9,25,26,29]. Misconceptions in mental models might lead users to behaviors that do not represent their actual needs [43].

Many existing studies investigated the mental models of individuals in the scope of cybersecurity in different technological contexts in the scopes such as encryption [28,48], threat perceptions of PC users [44,45], decentralized identity wallets [27], adversarial machine learning [7]. All studies highlight the importance and impact of mental models in (in)secure behavior.

Investigations of Privacy Mental Models. Mental models inform the behavior of users and profoundly contribute to adopting new technologies [2, 26, 41, 49]. Privacy mental models, in particular, have been shown to impact technology usage in the scope of the digitization [3,4,10,14,15,17,19,46,47,50,51].

Early investigations particularly considered internet usage and responses to threats [26] demonstrating that individuals with a better technological understanding perceive more risks, but do not necessarily take better precautions compared to individuals with a lower understanding.

Many investigations of IoT settings and smart homes in particular, showed that privacy concerns can form usage barriers that hinder people [2,41,49]. The specific concerns are rooted in the physical security of households or hackers gaining access to IoT devices [50, 53] calling out the need for the awareness of data collection, especially when data is sent through the internet [19,24,33,35].

2.3 Investigations of Digital Health Records

Several researchers investigated perceptions of patients and healthcare providers regarding EHRs and PEHRs. In terms of methods, either conversation-based interviews (cf. [30,31]), card-sorting exercises (cf. [13]), and online surveys [5] were predominantly used to investigate the patient's attitudes towards EHRs. Not surprisingly, patients consider ease-of-use as a dominant factor in adopting (P)EHRs [5,34,38].

The literature produced mixed results regarding privacy perceptions which might be linked to cultural differences. Privacy was frequently considered an important topic in the context of digital health in general [34]. When asked for specific concerns, study participants mentioned unauthorized access, e.g.,

in case PEHR data volumes get lost [30], misuse of the data stored in EHRs, e.g., exposing individuals with stigmatized diseases [31], or increased health insurance contributions for patients with treatment-intensive health conditions [22, 31]. This was shown in the countries Australia (EHR) [30], Canada (PEHR) [5, 22]. These findings related to health insurance perceptions further motivate our investigation because in Germany the health insurances serve as providers for the access apps, yet, should not be able to access all data. Privacy concerns were expressed in the context of hacking attacks [6], since hackers might attack EHRs to gain sensitive information which are concerns similar to those expressed in the context of IoT devices [50, 53]. American studies [13] showed that American patients want to have granular privacy control over how their health data is shared []. However, Swedish patients perceived that health care professionals having access to their EHR would be in the patients best interest and strict access guidelines instead of access control would be a sufficient security measure [30].

Further investigations, however, showed the opposite meaning that individuals might not have privacy concerns because they highly trust the central infrastructure and the Hippocratic Oath from doctors [30]. This was shown for Sweden [30] and Canada where patients stated to adopt the PHR based on trust in the PHR platform [5]. Finally, while Canadian patients were open for using the PEHR [5, 22], they were not aware of the PEHR already existing and being available [22]. The same was observed for Swedish patients [30].

Overall, this shows that different countries and cultures need to be investigated individually because privacy perceptions – much as the concept of privacy in itself – are highly individual.

2.4 Summary

There are many ways to realize EHRs. Germany uses a national infrastructure where health insurances serve as app providers. Yet, related work repeatedly showed that patients have different privacy concerns involving data access of health insurance companies. This paper investigates patient perceptions of this unique infrastructure by capturing mental models through drawing. This adds a new perspective to the literature because conversation-based interviews or surveys were predominantly used.

3 Method

To understand German citizens' existing mental models regarding EHRs, we conducted 21 semi-structured interviews. The interviews consisted of five parts as detailed below. One was a drawing exercise asking participants to sketch their expectations of the German infrastructure. We recorded all interviews with a microphone and a camera and took pictures of the drawings (cf. Appendix A.1 for the complete interview guide). On average each interview took 30 minutes, and participants were compensated with 10€ Amazon vouchers.

Table 1: Demographics of the sample.

ID	Age	Gender	Occupation	Highest Education
P1	18	m	School Student	Still at High School
P2	18	n/a	School Student	Still at High School
P3	18	f	School Student	Still at High School
P4	67	m	Retired	University
P5	26	w	Student	University
P6	27	m	Student	High School Diploma
P7	22	m	Student	High School Diploma
P8	27	f	n/a	University
P9	57	f	n/a	Apprenticeship
P10	23	m	n/a	Apprenticeship
P11	84	f	Retired Accountant	Apprenticeship
P12	89	f	Retired Pharmacist	University
P13	27	f	Physical Therapist	University
P14	54	f	Teacher	University
P15	52	f	Teacher	High School Diploma
P16	18	m	School Student	Still at High School
P17	53	f	Teacher	Apprenticeship
P18	55	m	Manager	University
P19	32	m	Engineer	University
P20	49	m	Firefighter	High School Diploma (FH)
P21	48	f	Sports Therapist	University

Participants & Recruitment. We recruited 21 participants by advertising through mailing lists, social networks, and word-of-mouth. All participants needed to be at least 18 years old. Further, they had to reside in Germany and currently actively use the German health care system by having at least a family doctor. Nine participants identified as male, eleven as female, and one preferred not to say[1]. The average age of the participants was 41.41 years ($min = 18$, $max = 89$, $SD = 21.82$).

The participants had diverse backgrounds with six being students at school or university. Three were retired. Four participants had a finished apprenticeship as highest education, nine had a university degree, four had a high school diploma and four were still in high school. Table 1 provides a detailed overview. All interviews were conducted in German.

Affinity for technology was assessed by the ATI scale which ranges from 1 to 6 [20]. Our sample had an average ATI score of 3.48 ($min = 2$, $max = 4$, $SD = 0.68$).

Study Procedure. The procedure of the semi-structured interviews was as follows:

1) Welcome & Consent: Before the interview, participants were informed about their rights, the captured data, and that they can abort the study at any time without negative consequences. They were further informed that the interview was audio-recorded and transcribed before analysis and that the drawing exercise was filmed without capturing their

[1]There were further answer options (i.e. prefer to self-describe, and diverse). Still, none of the participants chose them.

faces. This and further information were given to them on a participant information sheet that included a consent form that participants were asked to read and sign. Questions by participants were answered during this process.

2) Warm-Up: At the beginning of each interview, participants were asked questions about how their family doctor stores their patient data. Next, we asked about their knowledge and usage of the digital patient file. Afterward, they were given a printed information text on the *general idea* of EHRs in German, meaning that EHRs are stored electronically. Details of the infrastructure were not included in this part to not bias participants.

3) Drawing Exercise – Expectations: Based on the general information text participants read, they were asked to draw a model of the EHR based on their personal expectations. To make it easier for them to get started, they were handed prepared entity pictograms (i.e., doctor, health insurance, patient, generic server, generic devices, and patient file) as well as distractor pictograms next to a sheet of DIN-A4 paper and pencils in various colors as recommended by related work [33, 52]. Further, they were instructed to draw their model in the following scenario to make it easier for them: a doctor wants to access an existing EHR because this doctor is visited for the first time. To get a better understanding of the drawn models, participants were asked to think aloud [8] while drawing such that we could understand their thinking process. When asked to explain their thoughts on how entities are connected, some participants drew arrows. Additionally, they were asked follow-up questions after finishing their drawing to ensure all drawn parts were explained in detail.

4) Infrastructure Perceptions: To round off the interviews, the interviewees were then shown the real model of the German EHR infrastructure as seen in Fig. 1. The infrastructure was explained to them using the scenario of visiting a new doctor as detailed above. After the interviewer has made sure that the interviewee understood the model, the interviewee was asked follow-up questions specifically considering their perceptions of the real infrastructure. The interview was concluded by asking participants about ideas for improving infrastructure and whether this would change the interviewee's willingness to use EHRs. Before the interview ended, participants were given a chance to add any further comments or statements.

5) Demographics & Compensation: The recording then ended and participants were handed a tablet to answer a demographics questionnaire. As the last step, participants were compensated by an Amazon voucher with a value of 10€ (roughly 10 US dollars).

Data Analysis. Before the analysis, we first anonymized all captured data. Audio transcripts were transcribed into written form. To ensure participant identification is not possible by their handwriting in their drawn models, their writing was concealed by machine text. Next, two researchers independently analyzed the properties of the sketches by listing entities chosen and drawn by the participants, the purpose of the entity, and its communications. The researchers compared their lists and resolved disagreements in a meeting. Next, it was analyzed whether participants had functional or structural mental models [37]. For this, the transcripts were also considered to make sure that there are no misunderstandings.

The transcripts were also analyzed by thematic analysis [11]. In the first round, we conducted open coding by assigning codes to meaningful and relevant concepts focusing on our research questions. One researcher who was familiarized with the data generated an initial codebook. The codebook was then verified by a second researcher who was present during some of the interviewers and also had familiarized themself with the transcripts.

This codebook comprised 16 codes (see Table 2 in Appendix B). Following the methodology guidelines for conducting thematic analysis [11], one researcher applied the codebook to all statements. This was then verified by the second researcher and disagreements were resolved. Please note that guidelines for thematic analysis advise against double or multiple independent codings and using the inter-rater reliability to prove reliability [12, p.278-279]. This is because qualitative research acknowledges the influence of the researcher on the process [12]. Finally, four themes emerged from our analysis: (1) the role of health insurance companies, (2) the role of patients, (3) perceived risks, and (4) knowledge gaps and misconceptions.

Ethical Considerations. We took several precautions to protect the identities of our participants. Audio recordings were anonymized and transcribed before analysis. The participants' handwriting in the sketches was concealed by machine text. The consent form had detailed information about the data captured in the study, and the participants' rights and was compliant with the GDPR and national data protection laws. The consent form further mentioned that the study could be aborted without any negative consequences.

Since the study does neither have any risks beyond normal every day nor cause psychological harm, our institutions did not require formal IRB or ERB approval for the kind of study that we did. However, as stated above, we adhered to the strict (inter)national data protection laws and followed best practices for research conduct and transparency.

Limitations. Like most qualitative and exploratory investigations, our study is subject to several limitations that must be considered when reading our results.

First, our study is based on self-reported data, which might be biased due to social desirability, availability bias, and wrong recalls or self-assessments. Consequently, our data

only reflects the highly subjective views of our participants. Further, none of the participants had experience with using the German EHRs, hence their assessments are based on their expectations and might be different in case they actively use the EHRs. Yet, we wanted to capture patients' intuitive expectations of this new infrastructure since this also contributes to adoption. Still, especially in evaluating mental models, different subjective models of a concept like the digital health record are beneficial for research on how to counteract respective misconceptions or knowledge gaps. Having comparable experiences is quite challenging due to the high number of health insurance companies and the respective high number of EHR apps. Nevertheless, future work should investigate the actual usage of German EHRs.

Second, while we tried to recruit a diverse sample, our sample might not be representative of the entire German population. Still, our exploratory investigation served as a first step to investigating mental models of the Germans EHR infrastructure. Future work should investigate a representative sample.

Third, our investigation considers the specific use case of Germany and the German infrastructure. Based on that, our results regarding the perceptions are limited to the German system. However, the perceptions of the participants regarding the specific infrastructure can be used to inform the design of other infrastructures as well.

4 Results

The section details the results of the interview study and the drawing exercise. We first describe the sketches of the expected infrastructure. Next, we detail the results from the thematic analysis theme by theme. Whenever meaningful, we provide quotes from the participants that were translated from German. We further provide quantities of mentions to give the reader an impression of how often a specific aspect came up during the semi-structured interviews. However, this should not be mistaken as an attempt to quantify our results. RQ1 is answered in 4.1, 4.2, and 4.3. RQ2 is answered in 4.4, and RQ3 is answered in 4.5.

4.1 RQ1 - Overview of the Sketches

The sophistication of the participants' sketches was quite diverse. Fig. 2 provides some examples. None of the models intuitively described the German infrastructure correctly.

Functional vs Structural Models. Out of the 21 participants, only five participants (P4, P5, P7, P16, P18) had a structural mental model of the EHR which however had several parts that do not correspond to reality. The functional mental models were as their definition says quite simplistic. Most of them just listed a few entities with arrows between them. Some participants with functional models actively voiced knowledge gaps (see Sec. 4.4). A few with structural models (e.g., P5 in Fig. 2b) explained in detail how the data exchange and data management work in their understanding.

Entities. Intuitively, participants chose different entities to be part of their mental model. All participants included patients, doctors, and a central server. Most participants included the health insurance company in their sketches. However, all participants verbally mentioned that health insurance has a role in the infrastructure. Few participants also included other healthcare providers, such as emergency doctors or pharmacists because their access is needed in certain situations. Many participants integrated a smartphone to have data access. Two participants integrated a *"national authority"* (P3, P19). For a full overview of all entities drawn by individual participants, the reader is referred to Table 3 in Appendix C. This table also included access rights and data storage locations.

Similar results were found regarding mental models of decentralized identity wallets, where participants also considered different entities as part of their mental models, and reflected on the trustworthiness of these entities [27].

Storage Location. The participants sketched and mentioned several locations when it came where the EHRs are stored. Most participants mentioned *"a central server"* hosted by authorities, sample comments mentioning the central server are:

> P7: *"I want it [the EHRs] to be managed by a public authority, that would be my dream, like the public health authorities."*

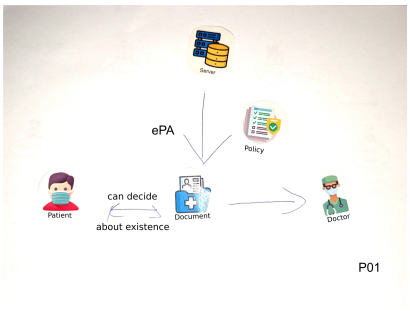
Further participants considered the *"health insurances"* (P8, P11, P16) to host the EHRs, e.g.:

> P11: *"The server is hosted by the health insurances because those should be trustworthy."*
>
> P8: *"By the health insurance, or somewhere else, I don't know where exactly."*

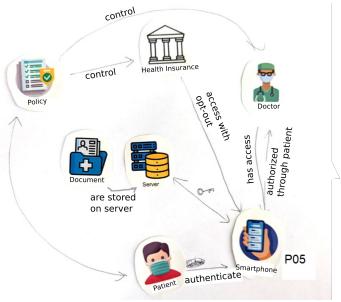
Two participants (P12, P14) considered the doctors to store the EHRs, while some participants considered several storage locations, instead of one, such as *"on a chip card*[2] *and a server"* (P18), or *"on a mobile device, the patient card, a server and with the health insurance"* (P6).

Access. Data access was an essential topic in most interviews. Considering the expected data access, the first group of participants expressed an *unspecified* access to EHRs by different entities like *"health insurances"* ($N = 5$), *"patients"* ($N = 4$), and *"doctors"* ($N = 3$). Some participants also mentioned *"hospitals"* ($N = 1$), *"pharmacists"* ($N = 1$) and

[2] By this, they mean the German health care chip card that each member of a statutory health insurance has.



(a) Functional Mental Model P1



(b) Structural Mental Model P5

Figure 2: Participants' sketches showing examples of their mental models. We replaced participants' handwriting with digital labels to enhance readability, and provide participant anonymity and for translation purposes.

individuals close to the patient, such as *"family members"* ($N = 1$) and those with access when the *"patient gave consent"* ($N = 1$). Yet, most participants expected doctors only to have access if the *"patient grants it"* ($N = 12$).

When asked about specific access rights, the patients considered different entities to have *complete read and write access* to their EHRs, namely health insurance companies ($N = 2$) and patients ($N = 4$). There were also various expectations towards doctors ($N = 4$), doctors in charge of the patient ($N = 1$), and doctors in emergency situations ($N = 1$).

Participants expressed various entities to have *complete read access* only to their EHRs, particularly health insurance companies ($N = 3$) and patients ($N = 1$).

Access Control. This brings us also to the topic of access control rights. Here participants had several ideas on who is managing access control to their EHRs.

Most participants considered the patients to be somehow involved in controlling access rights, such as:

P7: *"The best way, at least the way I hope, is that the patient first has to confirm somehow that he [the doctor] is allowed to do that [access the EHR]."*

P14: *"And [the doctor] should only have access to the server via the patient. So if the patient says, 'yes, okay' then the doctor can access it."*

Among them, some participants made a connection to existing granting of permissions in the medical context required by the GDPR, such as:

P8: *"I have to sign such documents all the time and consent that my data is shared. Because of that, I added that to my model out of habit."*

Yet, some participants also considered doctors and health insurance companies to have access control rights.

P11: *"Basically, I would say the family doctor. That's the one, right? But if he should give me the documents for the next doctor, he practically hands that over, does he? [...] In principle, the health insurance company is probably the highest, which also has a little bit of control over it."*

P16: *"And if one is then with the physician and the physician asks, 'I need or I would like to have access to provide treatment', then the physician can put a request to it, which must be confirmed by the patient and by the health insurance."*

Similar to our results access control played an important role in the mental models of decentralized identity wallets [27].

Delete. One participant (P19) mentioned that patients should be allowed to delete records.

4.2 RQ1 - Theme 1: The Role of Health Insurance Companies

The first theme identified in our analysis considers the specific role of health insurance companies within the infrastructure of EHRs specifically considering privacy and trust aspects.

Privacy towards Health Insurances. A slight majority of participants considered privacy towards health insurance as essential. Particularly, they were concerned that by having access to detailed patient data, health insurance companies might increase the contributions of patients with certain diseases or risks. Sample comments are:

P17: *"The health insurance and access to the data, I think that's kind of pretty difficult. [...] You might be categorized differently in terms of contributions, so at least when you apply for health insurance, I think that's very, very tricky if they had access to [the EHRs]."*

P18: *"The health insurance company needs no information at all, except about what is needed for billing. That's all they really need. It's nice when the health insurance company cannot access the content. In the best case, the information that the health insurance company needs*

is sent to them from the server. That's how it would be desirable from my point of view."

These results confirm concerns regarding the contributions to the health insurance companies from related work that investigated Australia [30] and Canada [5, 22].

Health Insurances as Trust Anchors and Backup. The remainder of the participants, however, voiced opposite opinions, specifically considering the health insurance companies as a trusted entity that also observes whether health care providers, such as doctors act genuinely:

P5: *"Access for the health insurance would probably be good, if that is somehow regulated in such a way that the health insurance automatically, if that is officially your health insurance, also always has access to your file, because I can imagine that for emergencies, if you yourself somehow don't have the possibility to authorize it, it's kind of stupid if your health insurance doesn't have access to it, besides, it feels like it has access to all things anyway, right?"*

P16: *"The server is provided by the health insurance because people trust it or at least it should be trustworthy."*

Trust playing a role in mental models was also observed in comparable studies [27].

Health Insurances as App Providers. As evidenced by the privacy issues regarding health insurance companies and the opposite opinions regarding trust anchors, it is challenging to provide a solution that fits the needs of all individuals. Some participants also commented on the health insurance companies as app providers. The first group considered this to be a negative aspect:

P19: *"If health insurances have access, they could change the contributions for each individual. I don't trust that all health insurance companies will be completely trustworthy 100% of the time. I don't trust that they wouldn't try to get some kind of benefit through the app provided by them."*

Several participants even suggested that the app provider should not be the health insurance companies:

P16: *"There should be a law, which describes the all guidelines exactly. Yet, health insurance companies might maybe find some loophole, you never know. So it's not that I suspect that they will do that [...] I trust the state more than the health insurance companies, which are still an institution somewhere, which also have to earn money."*

P20: *"Of course [the health insurance] needs some access. That is convenient, but I don't think that it should be involved in providing the app. I think that is very critical."*

P17: *"The health insurance company is only allowed to see prescriptions. I find that somehow strange that they then provide the app."*

4.3 RQ1 - Theme 2: The Role of Patients

The second theme revolves around the role of the patients within EHR infrastructures.

Control by Patients. First, participants expressed to have a variety of benefits of using the German EHRs specifically focused on the control exerted by patients:

P5: *"I really like that the patients can authorize doctors. They can also use the app to revoke that."*

P7: *"And then I can simply change policies by the app? That's really awesome."*

However, since patients using the EHR have quite a lot of responsibility, this might result in problems, because patients could be overwhelmed:

P8: *"On the other hand, it could also be too much for people who do not have an overview of what could be important or not. Where it is then perhaps easier to have all the findings and then can if they are generally not so familiar with digital media."*

P14: *"That is way too complicated for – and I'm not even careless – but there are people who think about it even less than I do, I think. So, it is much too complicated. In the end, the patient probably can do everything."*

Data Deletion by Patients. Patients are allowed to delete data from their EHRs. Some participants expressed concerns in connection with that, specifically fearing that people might "mindlessly" delete data that might, later on, be important or otherwise negative impacts on patient care:

P21: *"It might happen that important things are deleted, I mean information that is important for the doctor."*

P13: *"I just don't know whether it should be possible to delete these things as a patient, which I see a bit critically. I think it is important to archive such sensitive data."*

This risk is also reflected in official documentation provided to healthcare professionals specifically instructing them to make sure to have the data also locally stored at their practice [21].

Further participants feared that patients maliciously manipulate their files to gain a benefit:

P12: *"And that can't be right. There are professions where you need a health certificate. I needed one of those, for example. And if there are things in your patient file that contradict that, then you simply delete them. And apply for a job with a deleted patient file. That can't be good."*

P3: *"Patients should not delete information, they could repeatedly have prescriptions for prescription drugs written for them."*

However, data deletion by patients was not completely perceived in a negative way. Two participants liked that patients have the right to delete their data:

P16: *"Otherwise, I think it's good that the patient has so much control over the document. Because they have all the information, they can say whether they want the doctor to see the document, and then they can delete it if they no longer want to have it. I think that's good."*

P20: *"If something not needed is registered, then, of course, it will be deleted. This does not contribute to finding the truth about someone who is lying unconscious somewhere."*

4.4 RQ2 - Theme 3: Gaps & Misconceptions

As already stated above, the intuitive expectations of the infrastructure often did not match reality. Below, we detail knowledge gaps and misconception expressed by participants.

EHRs in Germany Are Not Available. None of our participants did use the EHR, although it was introduced in January 2021, which is more than one year before our study. The vast majority of them have not even heard about it. Here, several participants struggled to believe that the EHR is indeed available in Germany, leading to interesting conversations with the experimenter:

P20: *"I really like the idea and would welcome its introduction."*

P18: (after experimenter tells that the EHR is available for all patients in Germany) *"I just don't believe it."*

How Does Authentication & Authorization Work. Participants directly expressed several knowledge gaps they were aware of. This was in the context of authentication and authorization where participants struggled to explain how such a procedure might be done ...

P5: *"I know so little about it [health records], I don't know, for example, whether, well, because that must be password-protected somehow, or otherwise protected. [draws] I just noticed that I have a knowledge gap because I don't know where or how one could authorize the doctor."*

... but also in the context of consent, where participants expressed difficulty to explain when their consent is needed and when not, e.g.:

P9: *"That's kind of a big question mark for me because I don't know, I mean I know that somewhere there are agreements, a declaration of consent is made, also that others are allowed to access it. What exactly the guidelines are, I don't know."*

Doctors Have Full Control. Participants had various ideas on how doctors are involved in access control as mentioned above. In particular, doctors were given more power and more authority compared to reality:

P11: *"Basically, I would say right from the start, the GP [controls it]. But if he could give me the documents for the next doctor, he practically hands them over."*

P4: *"Patient don't have access to it. They just have to sign a consent form."*

One participant even said that doctors – once authorized – can also view the data offline:

P6: *"I mean, even if it [the health record] is not online, the doctor can still see the data if he wants to. That's why I would say that the doctor always has to sign a declaration of consent confirming that everything remains anonymous and is not passed on to third parties, that's what I would say. He can always call up the data."*

Emergency Access is Available. Two participants explicitly mentioned that emergency doctors have access to their data in case they are unconscious. However, the data can only be accessed in cooperation with the patients:

P4: *"If I have an accident, it would make sense if I wouldn't have to give them [the emergency doctors] any access authorization at all, but that they have access to my medical records via a special access right, so that they can act immediately."*

P7: *"So it would be cool if an emergency doctor could do that, for example, if they somehow arrive at an emergency scene or something and can then call up something in that direction. That would certainly be very practical."*

App Control is Impossible. When explained that the access control is done via an app or physically via the health insurance card, some participants still struggled with this:

P12: *"Using an app [to grant access], but the patient can't do that. What sense does that make? Nobody can grant access rights through a smartphone. No one is allowed to do that."*

Health Insurance Have All-Access. Most misconceptions expressed by participants were in connection to health insurance companies. Several participants thought that the health insurance has full access to their EHRs because this is needed for billing:

P11: *"In principle, the health insurance is the highest entity which also exerts most control."*

P12: *"The health insurance must be involved [in storing and managing the EHRs]. They have access to the data anyway through the prescriptions from the doctors."*

This seems to be related to a general misconception about the data that German health insurance companies can access that is already present with analogue patient records. As stated in Sec. 2.1 health insurances do not have access to patient

records. Perhaps interestingly, here misconceptions from the analog world diffused into the digital world.

Similar to the privacy aspects of health insurance companies, some patients saw full access as a requirement for health insurance companies to calculate their contributions:

P2: *"The health insurance companies also because they often want to know what kind of illness people have or whether people have any illnesses to somehow set the contributions accordingly or so that they can adjust to what will happen in the future."*

4.5 RQ3 – Theme 4: Perceived Risks

Besides the risks associated with the roles of the insurance companies and patients, participants perceived further risks based on the central storage of the EHRs and third parties.

Centralized Storage. Some participants considered the central storage of the German infrastructure to be problematic in the context of security:

P3: *"Therefore, my problem is not the handling with [the data], but rather that it is only stored in one place and that this is, so to speak, probably then quite or much easier to attack than paper files."*

Based on that, participants had various expectations and suggestions in the context of security, such as *"using encryption"* (P5) or *"something similar to two-factor authentication"* (P19), or *"letting doctors only access EHRs of patients who are physically present"* (P19).

Further, participants wanted that doctors have additional local files, such that too sensitive information does not get uploaded to a central entity:

P14: *"Overall, I like the idea of a general server, but I think that he [the doctor] should have a private file. Private, to store sensitive data. I don't want that everything I'm telling my doctor ends up on a server. No way!"*

Third Parties. Participants expressed that there might be further privacy risks if third parties, such as hospital providers, commercial operators, or the employer get access to the data:

P18: *"But these are also some purely commercially operated hospitals, and of course, you can't trust them, I say."*

P14: *"The health insurance company is a problem because if they know too much about a patient, which is already the case today, they may not accept him or her. Who knows what prejudices they may have? And also, I would like to say again here, the working world, employers, and so on, they should not be allowed to know everything either."*

Participants further praised that information about them might become more easily available when needed:

P14: *"Of course, I think it's great when I imagine I have an accident and then my name is entered and then everything that has been stored so far appears. And from this information, be it just my blood group, my life could be saved. I think that's great."*

P1: *"So I mean, you can perhaps also determine diseases that are perhaps somehow related, also sooner as a doctor alone."*

Finally, some participants liked they do not need to bring any existing doctor's letters or other kinds of documents with them and that burden is taken away from them:

P18: *"An advantage is that the patient does not have to bring anything."*

P5: *"I'm also like that, I tend to lose documents and then need to look for them forever. And I imagine that it's very practical to have them somehow online."*

Missing Assurance. Similar to other existing studies about data sharing in different domains [19, 24, 35], we found that participants want options for (a) consent sharing and (b) assurance about the status of their EHR:

P18: *"At the moment, the patient basically has hardly any control options. He can look at these documents and delete them, but he doesn't know whether the server always shows all the documents. [...] He probably has, I don't know, any knowledge about what is stored on the server and who has looked at it or so, so he is only at the end and he gets access to his data via an app. But who says that the data is displayed in full or that it then becomes transparent?"*

5 Discussion

In this paper, we explored the mental models of German EHRs, which is a national infrastructure that stores the health data of all German citizens wishing to use it. In the remainder, we discuss the management options of the German EHR by app and health card, perceptions of the role of health insurance companies, data manipulation, the involvement of patients, and their privacy implications. We further provide key takeaways and recommendations for digital infrastructure providers, e.g., developers and system designers, and healthcare providers. Finally, we compare our findings with results of similar studies conducted in other countries.

App- & Card-based Management. Patients can use an app from their health insurance company or an electronic health card to authorize access to documents in their EHR. While the electronic health card is an easy and existing way to authorize data access, it also comes with many limitations. Patients must be physically present at the doctor's office to manage their documents. Since doctor visits are already quite limited

in terms of duration, it is questionable whether doctors would indeed take more time to allow patients to browse and manage their data. While having card-based access is a possible fallback mechanism, it is unrealistic for actual usage.

Within the context of other IT systems in Germany, having such an app is quite a unique aspect. No participant initially thought that access is controlled this way. Having health insurance companies as app providers results in tension because (a) the companies only should have access to data needed for billing purposes, yet (b) the app allows patients also to add or delete documents. Consequently, patients need to trust that health insurance companies do not access this data.

The decentralized infrastructure was criticized by our participants for several reasons: trust assumptions are made towards health insurance companies, and there might be impacts on convenience or ease of use. Further, certain groups of individuals are excluded, development and maintenance come with challenges, and the apps create attack vectors. Finally, the current infrastructure did not match the participants' mental models which in turn might result in a low adoption as also shown in related domains, such as encryption [28, 48].

Takeaway 1: Challenges for Chip Card Users

Patients with the electronic health card only get a very limited service and are dependent on their doctors to exert control over their EHR.
Recommendation 1: The government should offer different possibilities for EHR access (e.g., smartphone & desktop apps, options for people without technical devices like kiosks). Patients should not depend on any healthcare provider to manage their EHRs.

Perceptions on Role of the Health Insurance Company. While most participants expressed concerns that insurance companies might use EHRs as part of their business model, some participants considered them to be a trusted supervisory authority that ensures the doctors do not break policies.

To dive into this more deeply, we need to understand the details about the German health insurance system given in Sec. 2.1. Further, we have to note that patients with statutory insurance always get the electronic health card which is needed to use EHRs. Private patients rarely get such a card meaning that the EHR is for patients with statutory insurance.

As stated above, private insurance companies rely on pre-existing conditions when making the decision to insure an individual or calculate contributions[3]. Consequently, it is already part of the business model. Current contribution models are not dynamic because private health insurance companies are only allowed to increase contributions if they can prove that the overall costs are rising. Hence, it is not allowed to consider new conditions. Since private insurance companies currently are not part of the EHR, private patients do not have to fear consequences like changing contributions. Further, it is not allowed by law. However, the benefits of EHRs should also be available to patients with private health insurance, and the infrastructure should consider that.

Health insurance companies need certain information about patients to deliver their service as rightfully assumed by our study participants. Yet, the information the statutory insurances receive is limited to that needed for the billing process [39]. The German infrastructure models this process, yet participants particularly struggled that health insurances provide the app to serve as a data controller. When installing the app, consent forms might ask patients to consent to share more data with health insurance companies than needed. Considering that most participants had a wrong mental model here, it might be easy for a health insurance company to get consent from their clients. Further, health insurances in Germany also have optional bonus programs rewarding patients for specific actions, e.g., yearly check-ups or being a sports club member. Currently, insurance apps promote these programs allowing them to combine even more data.

Besides the results from our study, this results in more problems: first, patients no longer having German health insurance might lose access to their EHRs which is difficult from the GDPR perspective. Second, the landscape of different apps is fragmented since 85 insurances offer EHR apps which also defeats the cost reduction efforts of EHRs. Third, the development, maintenance, and test process for the apps is challenging because patients might have issues with data that the health insurance is either legally not allowed to be seen or the patient does not want to share it, and fourth, as stated above individuals that do not want to or cannot use an app have challenges to overcome in case they wish to exert control over their data.

For the reasons above and a better alignment with patient expectations, there should be a central infrastructure. Access software, such as apps, should be provided by an independent provider, e.g., the one that also provides the server. Further, API documentation should be published in case individuals wish to use their own system to access their EHR. For this, a central access control system is needed that allows authenticating patients to prevent data leakage to others.

Takeaway 2: Insurances Should Not Provide Apps

Having health insurance companies as app providers is a challenging aspect that raises several concerns because there is tension between their responsibilities as app providers and patient privacy.
Recommendation 2: A central access and management option should be provided by an official entity different from the health insurance company.

[3] While it is legally challenging to refuse patients, the monthly contributions can get quite high, so some individuals choose statutory insurance in case of pre-existing conditions.

Data Manipulation. Some participants considered their insurance company a controlling entity that ensures doctors act genuinely. Currently, it is challenging to reveal error-prone data without the knowledge of a healthcare professional. Insurance companies cannot have this function either because of the trust aspects. Further, patients rightfully have a lot of power over their data. While most patients likely act genuinely, there might be cases where patients maliciously manipulate their files. Further, even if patients act genuinely, they might misjudge the importance of the stored data. Official documentation for doctors instructs them to make local copies of documents to ensure their availability. However, doctors are human as well and might forget this. A possible solution for that is allowing patients and healthcare professionals to mark certain records for verification by a trusted third entity that can trigger an investigation in case a record is suspicious.

Takeaway 3: Dispute-Resolution is Needed

Patients and doctors might add or remove data that is useful and needed.
Recommendation 3: Methods for record verification should be provided. Further, if access is granted to doctors, a local copy should be stored automatically.

Patient Involvement. Data access to the German EHR without patients is impossible. This gives patients the ultimate power over their data which was requested by patients studied in other countries [13]. Yet, as also commented by our participants, in health care, there might be situations where the patient is not available, but data access is critical, for instance, in case of an accident. Currently, there is no way for doctors to access the EHR. Similar to other scenarios with fallback access, there might be a scratch field on the patient's electronic health card that allows data access. A scratch field is a hidden field on the electronic health card similar to a scratch card. In an emergency, doctors could access the EHR by physically uncovering credentials under the scratch field to ensure the best possible patient treatment. Patient could see that someone uncovered the credentials, since removing a scratch field is irreversible.

Takeaway 4: Provide Emergency Access

The ultimate power of patients limits data access in an emergency situation.
Recommendation 4: Provide a secure and easy-to-use way for emergency data access.

German Perceptions vs Other Countries. This section compares our results to related studies conducted with Canadian [5, 22], American [13], and Swedish [30] patients.

Similar to our participants, for Canadian patients the biggest motivator for adopting PHRs is having access to their own medical data, and perceiving the PHR as useful in terms of usability, functionality and accessibility [5]. This is reflected in our findings by participants liking the extend of control patients have over their EHR and also confirms the findings of the American studies [13]. Our participants perceived the necessity of trusting health insurance companies as critical, since they function as providers of the different German EHR applications. This contrasts results from Canada [5].

Perhaps interestingly, since EHRs are quite new, Canadian [22] and Swedish participants were not aware of their digital infrastructure similar to our participants [30]. This implies that right now participants do not use the EHR. Resulting in their mental models not having impact on their behavior. Since we evaluated some of their misconceptions, future research can develop measures to counteract those and therefore get patients to use the EHR.

The similarities in our results we found compared to studies from other countries, show that our four takeaways based on patients mental models of the German EHR can also be applied in a broader context. However, we would like to add that the specific cultural context also should be carefully considered when designing EHR infrastructures.

6 Conclusion & Future Work

This paper investigated mental models of electronic health records (EHRs) of German citizens in the context of introducing nationwide centralized EHRs. In this investigation, we focused on aspects related to data sharing, privacy, access management, and trust. We interviewed 21 individuals that currently reside in Germany and use the German health system. Using semi-structured interviews and a drawing exercise, we captured the mental models of patients identifying four core themes.Mostly, participants had incorrect ideas regarding the role of health insurance companies. In Germany, they can only access the data needed for billing purposes, yet participants thought that health insurances have all access, manage the EHRs, or even act as a trusted authority. In the German EHR system, health insurances serve as app providers. The apps can be used by patients for policy management of their EHR. This results in tension between the insurance company as an app provider and the fact that they only have limited data access. Further, patients in Germany are allowed to add and delete EHR documents. This was critically questioned by many participants who feared a negative impact on diagnoses.

Based on our investigation, we provide valuable insights in the form of recommendations for digital infrastructure providers, such as developers, system designers, and healthcare providers. Future work should specifically investigate possibilities for dispute resolution, e.g., in case a patient or doctor adds non-credible data. Further, mental models of other types of infrastructures should be captured and compared with the German ones.

Acknowledgements

This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC 2092 CASA - 390781972. It was furthermore supported by the German Federal Ministry of Education and Research (BMBF) in the project "Verbundprojekt: Digitale Fitness für Bürgerinnen und Bürger – realistische Risikowahrnehmung, sichere Routinen - DigiFit" ("Joint project: digital fitness for citizens - realistic risk perception, secure routines - DigiFit") – grant number 16KIS1646K.

References

[1] Drei viertel der deutschen wollen elektronische patientenakte nutzen. *Bitkom e.V.*, Mo., 06.12.2021 - 10:10.

[2] Noura Abdi, Kopo M. Ramokapane, and Jose M. Such. More than smart speakers: Security and privacy perceptions of smart home personal assistants. In *Proceedings of the Symposium on Usable Privacy and Security*, SOUPS '19, pages 1–16, Berkeley, CA, USA, 2019. USENIX Association.

[3] Tousif Ahmed, Roberto Hoyle, Patrick Shaffer, Kay Connelly, David Crandall, and Apu Kapadia. Understanding physical safety, security, and privacy concerns of people with visual impairments. *IEEE Internet Computing*, 21(3):56–63, May/June 2017.

[4] Noah Apthorpe, Yan Shvartzshnaider, Arunesh Mathur, Dillon Reisman, and Nick Feamster. Discovering smart home internet of things privacy norms using contextual integrity. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(2):59, 2018.

[5] Norm Archer and Mihail Cocosila. Canadian patient perceptions of electronic personal health records: An empirical investigation. *Communications of the Association for Information Systems*, 34(1):20, 2014.

[6] Sahil Bhagat, D Fontaine, and K Gibson. Danish healthcare information technology-an analytical study of consumer issues. *Worcester, MA: Worcester Polytechnic Institute*, 2010.

[7] Lukas Bieringer, Kathrin Grosse, Michael Backes, Battista Biggio, and Katharina Krombholz. Industrial practitioners' mental models of adversarial machine learning. In *Eighteenth Symposium on Usable Privacy and Security (SOUPS 2022)*, pages 97–116. Usenix Association, 2022.

[8] Ted Boren and Judith Ramey. Thinking aloud: Reconciling theory and practice. *IEEE transactions on professional communication*, 43(3):261–278, 2000.

[9] Christine L. Borgman. The user's mental model of an information retrieval system: An experiment on a prototype online catalog. *International Journal of Man-Machine Studies*, 24(1):47–64, 1986.

[10] Denys Brand, Florence D. DiGennaro Reed, Mariah D. Morley, Tyler G. Erath, and Matthew D. Novak. A survey assessing privacy concerns of smart-home services provided to individuals with disabilities. *Behavior Analysis in Practice*, 13:11–21, 2019.

[11] Virginia Braun and Victoria Clarke. Thematic analysis. 2012.

[12] Virginia Braun and Victoria Clarke. *Successful qualitative research: A practical guide for beginners*. SAGE Publications, London, 2013.

[13] Kelly Caine and Rima Hanania. Patients want granular privacy control over health information in electronic medical records. *Journal of the American Medical Informatics Association*, 20(1):7–15, 2013.

[14] Eun Kyoung Choe, Sunny Consolvo, Jaeyeon Jung, Beverly Harrison, Shwetak N. Patel, and Julie A. Kientz. Investigating receptiveness to sensing and inference in the home using sensor proxies. In *Proceedings of the Conference on Ubiquitous Computing*, UbiComp '12, pages 61–70, New York, NY, USA, 2012. ACM.

[15] Hyunji Chung, Michaela Iorga, Jeffrey Voas, and Sangjin Lee. Alexa, can i trust you? *Computer*, 50(9):100–104, 2017.

[16] Donald P Connelly, Young-Taek Park, Jing Du, Nawanan Theera-Ampornpunt, Bradley D Gordon, Barry A Bershow, Raymond A Gensinger Jr, Michael Shrift, Daniel T Routhe, and Stuart M Speedie. The impact of electronic health records on care of heart failure patients in the emergency room. *Journal of the American Medical Informatics Association*, 19(3):334–340, 2012.

[17] Kovila PL Coopamootoo and Thomas Groß. Mental models: an approach to identify privacy concern and behavior. In *Symposium on Usable Privacy and Security (SOUPS)*, pages 9–11, 2014.

[18] B Devkota and A Devkota. Electronic health records: advantages of use and barriers to adoption. *Health Renaissance*, 11(3):181–184, 2013.

[19] Pardis Emami-Naeini, Sruti Bhagavatula, Hana Habib, Martin Degeling, Lujo Bauer, Lorrie Cranor, and Norman Sadeh. Privacy expectations and preferences in an iot world. In *Proceedings of the Symposium on Usable Privacy and Security*, SOUPS '17, pages 399–412, Berkeley, CA, USA, 2017. USENIX Association.

[20] Thomas Franke, Christiane Attig, and Daniel Wessel. A personal resource for technology interaction: Development and validation of the affinity for technology interaction (ati) scale. *International Journal of Human-Computer Interaction*, 35(6):456–467, 2019.

[21] Bundesminsterium für Gesundheit. Die elektronische patientenakte (ePA). https://www.bundesgesundheitsministerium.de/elektronische-patientenakte.html, 2021.

[22] Marie-Pierre Gagnon, Julie Payne-Gagnon, Erik Breton, Jean-Paul Fortin, Lara Khoury, Lisa Dolovich, David Price, David Wiljer, Gillian Bartlett, and Norman Archer. Adoption of electronic personal health records in canada: perceptions of stakeholders. *International journal of health policy and management*, 5(7):425, 2016.

[23] gematik GmbH. Die epa-app die angebote der gesetzlichen krankenkassen. https://www.gematik.de/anwendungen/e-patientenakte/epa-app, last-accessed 13-Feb-2023, 2023.

[24] Timo Jakobi, Corinna Ogonowski, Nico Castelli, Gunnar Stevens, and Volker Wulf. The catch(es) with smart home: Experiences of a living lab field study. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '17, pages 1620–1633, New York, NY, USA, 2017. ACM.

[25] Philip N. Johnson-Laird. *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Number 6. Harvard University Press, Cambridge, MA, USA, 1983.

[26] Ruogu Kang, Laura Dabbish, Nathaniel Fruchter, and Sara Kiesler. "my data just goes everywhere:" user mental models of the internet and implications for privacy and security. In *Proceedings of the Symposium on Usable Privacy and Security*, SOUPS '15, pages 39–52, Berkeley, CA, USA, 2015. USENIX Association.

[27] Maina Korir, Simon Parkin, and Paul Dunphy. An empirical study of a decentralized identity wallet: Usability, security, and perspectives on user control. In *Eighteenth Symposium on Usable Privacy and Security (SOUPS 2022)*, pages 195–211, Boston, MA, August 2022. USENIX Association.

[28] Katharina Krombholz, Karoline Busse, Katharina Pfeffer, Matthew Smith, and Emanuel Von Zezschwitz. " if https were secure, i wouldn't need 2fa"-end user and administrator mental models of https. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 246–263. IEEE, 2019.

[29] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. Too much, too little, or just right? ways explanations impact end users' mental models. In *Proceedings of the IEEE Symposium on Visual Languages and Human Centric Computing*, VL/HCC '13, pages 3–10, Piscataway, NJ, USA, Sep. 2013. IEEE.

[30] Elin C Lehnbom, Andrew J McLachlan, and Jo-anne E Brien. A qualitative study of swedes' opinions about shared electronic health records. In *MEDINFO 2013*, pages 3–7. IOS Press, 2013.

[31] Elin C Lehnbom, Andrew J McLachlan, and E Brien Jo-anne. A qualitative study of australians' opinions about personally controlled electronic health records. In *HIC*, pages 105–110, 2012.

[32] CY Lu and E Roughead. Determinants of patient-reported medication errors: a comparison among seven countries. *International journal of clinical practice*, 65(7):733–740, 2011.

[33] Karola Marky, Sarah Prange, Max Mühlhäuser, and Florian Alt. Roles matter! understanding differences in the privacy mental models of smart home visitors and residents. In Adalberto L. Simeone, Raf Ramakers, and Cristina Gena, editors, *20th International Conference on Mobile and Ubiquitous Multimedia*, pages 108–122, New York, NY, USA, 2021. ACM.

[34] Neethu Mathai, Tanya McGill, and Danny Toohey. Factors influencing consumer adoption of electronic health records. *Journal of Computer Information Systems*, 62(2):267–277, 2022.

[35] Mateusz Mikusz, Steven Houben, Nigel Davies, Klaus Moessner, and Marc Langheinrich. Raising awareness of iot sensor deployments. In *Proceedings of the Living in the Internet of Things: Cybersecurity of the IoT*, London, UK, 2018. IET.

[36] NHS. How to get your medical records. https://www.nhs.uk/using-the-nhs/about-the-nhs/how-to-get-your-medical-records/, 2021.

[37] Donald A. Norman. Some observations on mental models. In *Mental Models*, pages 15–22. Psychology Press, 2014.

[38] Jamil Razmak and Charles Bélanger. Using the technology acceptance model to predict patient attitude toward personal health records in regional communities. *Information Technology & People*, 31(2):306–326, 2018.

[39] Sozialgesetzbuch (SGB) Fünftes Buch (V) – Gesetzliche Krankenversicherung. Artikel 1 des gesetzes v. 20. dezember 1988, bgbl. i s. 2477, 1988. https://www.gesetze-im-internet.de/sgb_5/.

[40] Stiftung Gesundheitswissen. Die elektronische patientenakte (epa): Wie sie funktioniert und was sie bringt. https://www.stiftung-gesundheitswissen.de/gesundes-leben/e-health-trends/die-elektronische-patientenakte-epa-wie-sie-funktioniert-und-was-sie, 31.01.2023.

[41] Madiha Tabassum, Tomasz Kosinski, and Heather Richter Lipford. I don't own the data": End user perceptions of smart home device data practices and risks. In *Proceedings of the Fifteenth Symposium on Usable Privacy and Security*, SOUPS, Berkeley, CA, USA, 2019. USENIX Association.

[42] Ahmad Tubaishat. The effect of electronic health records on patient safety: A qualitative exploratory study. *Informatics for Health and Social Care*, 44(1):79–91, 2019.

[43] Joe Tullio, Anind K. Dey, Jason Chalecki, and James Fogarty. How it works: A field study of non-technical users interacting with an intelligent system. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '07, page 31–40, New York, NY, USA, 2007. Association for Computing Machinery.

[44] Rick Wash. Folk models of home computer security. In *Proceedings of the Sixth Symposium on Usable Privacy and Security*, SOUPS '10, New York, NY, USA, 2010. Association for Computing Machinery.

[45] Rick Wash and Emilee Rader. Too much knowledge? security beliefs and protective behaviors among united states internet users. In *Eleventh Symposium On Usable Privacy and Security (SOUPS 2015)*, pages 309–325, Ottawa, July 2015. USENIX Association.

[46] Charlie Wilson, Tom Hargreaves, and Richard Hauxwell-Baldwin. Benefits and risks of smart home technologies. *Energy Policy*, 103:72–83, 2017.

[47] Peter Worthy, Ben Matthews, and Stephen Viller. Trust me: Doubts and concerns living with the internet of things. In *Proceedings of the ACM Conference on Designing Interactive Systems*, DIS '16, pages 427–434, New York, NY, USA, 2016. ACM.

[48] Justin Wu and Daniel Zappala. When is a tree really a truck? exploring mental models of encryption. In *Symposium on Usable Privacy and Security*, pages 395–409. Usenix Association, 2018.

[49] Yaxing Yao, Justin Reed Basdeo, Smirity Kaushik, and Yang Wang. Defending my castle: A co-design study of privacy mechanisms for smart homes. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–12, New York, NY, USA, 2019. ACM.

[50] Eric Zeng, Shrirang Mare, and Franziska Roesner. End user security & privacy concerns with smart homes. In *Proceedings of the Symposium on Usable Privacy and Security*, SOUPS '17, pages 65–80, Berkeley, CA, USA, 2017. USENIX Association.

[51] Yu Zhai, Yan Liu, Minghao Yang, Feiyuan Long, and Johanna Virkki. A survey study of the usefulness and concerns about smart home applications from the human perspective. *Open Journal of Social Sciences*, 2(11):119, 2014.

[52] Verena Zimmermann, Merve Bennighof, Miriam Edel, Oliver Hofmann, Judith Jung, and Melina von Wick. 'home, smart home'–exploring end users' mental models of smart homes. In *Mensch und Computer 2018-Workshopband*, pages 407–417, Bonn, Germany, 2018. Gesellschaft für Informatik e.V.

[53] Verena Zimmermann, Paul Gerber, Karola Marky, Leon Böck, and Florian Kirchbuchner. Assessing users' privacy and security concerns of smart home technologies. *i-com*, 18(3):197–216, 2019.

A Study Materials

A.1 Interview Guide

This section provides the interview script using for the semi-structured interviews.

- **Welcoming & Consent** (not recorded)
  - *Welcome to this interview and thank you very much for participating. The interview will start with an introduction, where you will be asked some general questions about the topic and you will get some information about it. Then, follows a part in which I ask you to draw something here on the paper. Further, I'll ask you some questions about your drawing. At the very end, I'll ask you to fill in a short questionnaire about yourself. There are no right or wrong answers, I'm always interested in your personal opinion.*
  - *Please read this information sheet completely and sign it. If anything is not clear, please ask me.*
  - *Once, you're ready, I'll start the recording and let you know.*
- **Warm-Up** (audio-recorded)
  - *I'm starting the audio recording. Do you agree being recorded?*
  - *To get us started with the topic, I would like know: Do you know how your primary care physician stores your patient data?* (Possible help, if participant struggles: *Does they use paper files or maybe a PC or something else?*
  - *The main topic of today's interview are digital health records (German: elektronische Patientenakte, short: ePA). Have you heard about the digital patient file – short ePA – in Germany?*
    - (If yes:) *Can you briefly describe what it is?*
  - *Have you ever used any kind of digital patient file?*
    - (If yes:) *Do you use it regularly or just once?*
  - The participant gets the following information text about the ePA and is asked to read it: The electronic patient file (ePA) stores all important information on a patient's state of health and medical history. The idea is that data, such as medications taken, previous treatments or the results of imaging procedures are always available when a patient visits a doctor. Unnecessary multiple examinations and duplicate treatments can thus be avoided. Possible interactions between different medications can also be better taken into account in advance. The bundling of health-related information in the ePA is also expected to improve care in general. For example, the maternity passport, the yellow examination booklet for children, and the vaccination record will be available digitally from 2022. The most important medical data will be stored regardless of location and can be accessed from anywhere. The text is taken from official online documentation [40].
  - *Do you have any questions regarding the digital patient file?* (Questions about technical details were postponed to after the interview to not bias participants.)
- **Drawing Exercise** (audio- and video-recorded)
  - *Now, we start with the second part of the interview. Now, I'm interested in your idea, how the ePA works. For this, I'm asking you to make sketch of that using the paper and pens in front of you. We also have a few icons you can optionally use to make drawing easier for you, but this is optional. As explained before the interview, the drawing will be filled but your face cannot be seen. To make it a bit easier for you, we consider the following scenario: Assume a patient visits a new doctor who wishes to access an existing patient record. Do you have any questions regarding that?*
  - *While drawing please think aloud and explain me what you draw and why.* The experimenter asks questions about the entities and data drawn by the participants to get a full understand of the participant's ideas, such as:
    - *What happens with files?*
    - *Where are files stored?*
    - *Who has access to the files?*
    - *Who is allowed to manage the files?*
  - *Thanks for explaining everything to me.*

- **Infrastructure Perceptions** (audio- and video-recorded)
  - The participant is shown and explained the real model of the digital patient file based on Fig. 1:
    In this part of the interview, we take a look at the real German infrastructure using the scenario from before. The doctor requests a document from the ePA from a central server via their patient management system. All patients' ePAs are stored on this central server. There is one authorization database that defines the authorizations for different actors on different types of documents. There is also a policy for each document, on which individual access permissions are defined. Based on these, the server checks the authorization of the request and, if necessary, sends the requested document to the doctor. Patients use a mobile app from their health insurance company to create and edit authorizations for their documents. They can also use the app to view and delete all documents. Hence, the app changes the policy stored on the server accordingly or deletes the documents. Do you have any questions about that?
  - *Would you like to use this infrastructure?*
    - (If yes:) *Why?*
    - (If no:) *Why not?*
  - *Let's look at this infrastructure together. If you could decide yourself about the structure, access control management, etc., is there anything that you would like to change? If so, why? You can also just draw these changes on the paper. How does this influence your willingness to use the ePA?*
  - *Do you have further comments or questions?*
  - *I'm stopping the recording.*
- **Demographics & Compensation** (not recorded): Participants are asked to fill in the demographics questionnaire and the reimbursement form.

B Codebook

This section provides the codebook used to analyze the transcripts.

Table 2: Table displaying the qualitative codebook.

Code	Description	#
data_exchange	Participant explains how and between whom data is exchanged in the participants model	4
disapproval	Participant explains reason to refuse using the EHR	4
data_storage	Where the EHR is stored in the participants model	34
privacy_awareness	The participant wants to explicitly protect specific data	15
expectations	Requirements the participant expects from the infrastructure	22
access_rights_allocation	How access to the EHR is granted to different stakeholders in the participants model	43
terminological_misunderstanding	The participant had a different understanding of a specific term	25
knowledge_gap	The participants claims to not know something	22
editing_ability	Who can change the information in the EHR in the participants model	41
perceived_advantage	The participant perceives a feature of the EHR as an advantage	12
modification_idea	The participants has an idea for improving the real model of the EHR	26
positive_attitude	The participant has positive feeling towards using the EHR	30
skeptical_attitude	The participant is skeptical about aspects if the EHR	17
access_rights	The participant talks about who may access the EHR	84
access_method	The participant describes methods to access the EHR	23
risks	The participant perceives something as critical	31

C Additional Results

Table 3: Table displaying the entities of participants mental models, their access right, and the storage location of the EHRs. 👁 denotes read access, ✎ denotes write access, and 🏋 denotes access management.

ID	Entities	Access to EHR	Storage Location
P1	patient, doctors, server, health insurance	doctors, health insurance	server
P2	patient, doctors, server, health insurance, smartphone	patient 👁✎, doctors if granted, health insurance if granted	server
P3	patient, doctors, server, health insurance, national authority	doctors if granted	server of health insurance
P4	patient, doctors, server, health insurance, smartphone, emergency doctors	doctors 👁✎, emergency doctors 👁✎	server
P5	patient, doctors, server, health insurance, smartphone	doctors if granted, health insurance	server
P6	patient, doctors, server, health insurance, smartphone	doctors if granted	server of health insurance, chipcard, smartphone
P7	patient, doctors, server, health insurance, smartphone	doctors if granted, family, emergency doctor, family	server
P8	patient, doctors, server, health insurance	patient, doctors if granted	server of health insurance
P9	patient, doctors, server, health insurance, smartphone, hospital	doctors if granted, hospital if granted, health insurance	server
P10	patient, doctors, server, health insurance	doctors if granted	server
P11	patient, doctors (in charge), server, health insurance, smartphone, pharmacist	patient, doctors in charge 👁✎, pharmacist	server of health insurance
P12	patient, doctors, server, health insurance	doctors if granted, health insurance	doctors
P13	patient, doctors, server, health insurance, smartphone	patient, doctors, health insurance	server, App
P14	patient, doctors, server, health insurance	doctors if granted	doctors
P15	patient, doctors, server, health insurance, smartphone	patient 👁✎, doctors 👁✎, health insurance 👁✎	server
P16	patient, doctors, server, health insurance, smartphone, national authority	patient 👁✎, doctors if granted, health insurance	server of health insurance
P17	patient, doctors, server, health insurance, smartphone	doctors if granted	server
P18	patient, doctors, server, health insurance, smartphone	doctors 👁✎, health insurance 👁✎	server, chipcard
P19	patient, doctors, server, health insurance, independent national authority	patient 👁✎, doctors 👁✎	server of independent national authority
P20	patient, doctors, server, health insurance, smartphone, hospitals	doctors, hospitals	server
P21	patient, doctors, server, health insurance, smartphone	patient, doctors if granted	server

"Nobody's Happy": Design Insights from Privacy-Conscious Smart Home Power Users on Enhancing Data Transparency, Visibility, and Control

Sunyup Park, *University of Maryland, College Park*
Anna Lenhart, *University of Maryland, College Park*
Michael Zimmer, *Marquette University*
Jessica Vitak, *University of Maryland, College Park*

Abstract

As smart home technologies continue to grow in popularity and diversity, they raise important questions regarding ways to increase awareness about data collection practices and empower users to better manage data flows. In this paper, we share insights from 32 privacy-conscious smart home power users—individuals who have invested significant time, money, and technological prowess in customizing their smart home setup to maximize utility and meet privacy and security needs. We explore the drawbacks and limitations power users experience when balancing privacy goals with interoperability, customizability, and usability considerations, and we detail their design ideas to enhance and extend data transparency, visibility, and control. We conclude by discussing the importance of designing smart home technologies that both address these considerations and empower a wide range of users to make more informed decisions about whether and how to implement smart technologies in their homes, as well as the wider need for greater regulation of technologies that collect significant user data.

1. Introduction

Smart home devices (SHDs) have gained popularity in recent years, offering a convenient and efficient way to control and monitor various aspects of one's home remotely. SHDs range from smart speakers and thermostats to security systems and lighting, and they can be integrated with other smart devices, hubs, and apps to create a fully automated smart home environment.

While SHDs offer significant benefits—ranging from convenience and cost efficiency to added security and accessibility—they have also led to increased data privacy risks. In particular, researchers and privacy advocates have raised questions regarding the vast amounts of data devices collect about users and their environments, often without their full knowledge or consent [2,54]. This data includes information about household members' activities, movements, habits, and preferences, which can potentially be misused by hackers, manufacturers, government agencies, or others [31].

Ensuring data privacy and security requires continuous vigilance from consumers to be aware of the data flows across and between devices and the privacy policies and practices of the companies they purchase SHDs from, and to take necessary steps to secure smart home environments. Beyond these factors, research highlights that creating holistic approaches to protecting the privacy of smart home environments that address the different platforms, end-users, and data flows is very challenging [9].

In seeking to address common privacy concerns, prior research has largely evaluated the privacy attitudes and behaviors of "average" or "everyday" smart home users. However, power users—those who "use the devices more innovatively, efficiently, and thoroughly than ordinary users" [57] (p. 1743)—engage in a wider range of practices to maximize the utility of SHDs and implement strategies to track and manage data flows beyond what devices natively provide. Because this population is heavily engaged in researching device options and spending time and energy optimizing setup to balance functionality and privacy, "privacy-conscious power users (PCPUs)" are uniquely positioned to provide feedback and insights that everyday users may not consider.

In this paper, we share insights from focus groups with 32 privacy-conscious smart home power users to better understand the limitations of current smart home options and identify key areas for improving data access and control. By evaluating the limitations of current technologies and eliciting their design ideas for improving or enhancing data management and control, our focus on PCPUs provides a unique perspective on how to better design smart home technologies to match the needs of a full range of users. Thus, we focus our analysis on two research questions:

RQ1: What drawbacks and limitations do smart home power users identify in their current smart home setup?

RQ2: What design features do smart home power users want to see developed or expanded in future tools and platforms to enhance data management and control?

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.

USENIX Symposium on Usable Privacy and Security (SOUPS) 2023. August 6 -- 8, 2023, Anaheim, CA, USA.

.

Our findings highlight major tensions between the privacy and security goals our participants have and the customizability and interoperability of devices and hubs. Even though our participants spent significant time, energy, and money on their smart home setups, they found themselves sometimes limited in their options and having to make tradeoffs between privacy and functionality. Participants also described challenges managing devices with multiple users, whether other household members or visitors.

To address these challenges, our participants suggested three core areas for design improvements: data transparency, visibility, and control. Importantly, these power users stressed the need for interfaces and device management controls that accommodate a range of skill levels and privacy preferences, recognizing that most users lack the technical skills or desire to use advanced network management features or customizations to protect their privacy. At the same time, our participants also wanted advanced features, customizations, and data visualization options to accommodate their goals as power users. We conclude by reflecting on the barriers to accommodating these ideas and building tools that can provide control and privacy enhancement for a wide range of user types, and we consider how emerging standards and legislation might help in promoting data rights.

2. Related Work

2.1. Privacy Risks with Smart Home Technology

There are three major categories of privacy risks associated with SHDs. First, we consider the *location* of devices within the home. Consumers have significant concerns about data being collected in their homes [5,12,17,28,35] and find it more sensitive than data collected in public spaces [17]. Privacy concerns vary by room, with bedrooms [10,12], bathrooms [10,29,35], and children's spaces [51] being among the most concerning.

Second, people's concerns vary based on the *type* of data being collected. For example, audio data is frequently captured from smart speakers and TVs. Dunbar et al. [14] describe three categories of attitudes toward audio data collection: pragmatists, who have few concerns, are willing to trade privacy for benefits, and generally trust companies; guardians, who attempt to minimize data collection and safeguard data that is collected; and cynics, who rarely change default settings and lack a clear understanding of when data is collected and how it is processed. Video data, such as that collected from smart cameras in and around the home, also raises concerns [10,17,51], while research suggests smart home users have few concerns about raw data from more simple sensors (e.g., temperature, light) [25].

Third, researchers have found that *who* is collecting data is important to consumers [5,27,28]. Many consumers consider the reputation of technology companies when making decisions to purchase IoT devices [32,54], and users generally express trust that device manufacturers will collect data for legitimate purposes [26,56].

2.2. Privacy Design Work on IoT & Smart Homes

Usable privacy and security researchers have developed and evaluated various mechanisms to mitigate users' privacy concerns associated with IoT and SHDs. Perhaps the best-known design is privacy nutrition labels, which seek to increase transparency and support informed decision-making through standardized information about data collection and use. This work was initially carried out by Kelley et al. [22], before also being implemented by Apple in 2020 to provide standardized privacy labels for apps in the AppStore [39]. Emami-Naeini et al. [16] extended these labels to IoT devices, providing two layers of information to consumers: a more general label on the device packaging that contains information about security mechanisms, and a second label (accessible online) containing more detailed information about data collection and use.

The deployment of privacy nutrition labels has, however, hit roadblocks. There remains little incentive for developers of older apps to create or update privacy labels, and privacy labels themselves seem to be rarely updated once created [30]. Furthermore, even when developers think privacy labels are a positive thing, they are faced with challenges in creating them because of misunderstandings and the overall complexity of the task [30].

The IoT privacy and security community has also created tools aimed at increasing the visibility of data flows and provide users with additional privacy controls [47]. Tools such as IoTSense [7] and IoTSentinel [36] identify devices by analyzing network traffic, while IoT Inspector provides visualizations of device activity and traffic destinations [20]. Others have designed tools to increase the legibility of information flows and improve interpretability by providing users with more details about their connected devices and providing actionable choices [43,47]. Zeng and Roesner [55] designed a tool that includes location-based access controls as well as supervisory access controls to allow multiple users within a smart home to control their own privacy settings.

Information about data practices and data flows can be complemented with privacy notices as a part of privacy awareness mechanisms in smart homes. Privacy notices in the smart home context heavily focus on the use of visual and audio indicators on SHDs. For example, Song et al. [49] found that users were most interested in knowing the physical location of cameras and voice assistants, and preferred locator mechanisms that integrated visual and audio cues as the number of devices increased. In fact, many smart speakers now use lights to communicate the device's status to

users. However, Thakkar et al. [52] note that privacy notices in smart homes have a long way to go; current notices focus on individual stakeholders and devices, whereas the future of smart homes will inevitably consist of multiple stakeholders and devices.

2.3. Designing Smart Home Privacy Tools for Diverse Users

Researchers have also stressed the need to account for a diverse user base when designing SHDs and controls. For example, Chhetri and Motti [11] note that most SHDs lack user-friendly privacy controls, and they develop a framework to guide developers in building user-friendly privacy controls. In a similar vein, Kim et al. [23] created personas to help designers better understand potential users vulnerable to cybersecurity risks. Researchers have also identified several categories of design features to mitigate privacy harms, including transparency, privacy and security controls, and assistance for users [19,53]. In summary, features and diagnostic tools should be simple, proactive, preventative, and provide users with transparency and control [6,21].

Researchers have also considered the privacy needs of bystanders, including non-primary users as well as those whose data is captured from incidental interaction with devices. Yao and colleagues [53] found that bystanders' privacy concerns are more contextually dependent than users; bystanders wanted to cooperate with smart-home owners to negotiate privacy needs. Ahmad et al. [3] argue that devices should be designed to afford users and bystanders "tangible privacy" through features like camera covers, physical on/off buttons, and clear on/off indicators. On the other hand, Thakkar et al. [52] found that bystanders might not know or neglect users' privacy concerns when privacy awareness mechanisms are implicated; they suggest having a separate bystander mode within device controls. Similarly, Marky et al. [34] found that visitors' lack of awareness limits their ability to protect their privacy.

As a whole, research to date has identified several privacy risks associated with data collected by SHDs and frameworks for design tools to mitigate those risks. The present study extends this prior work by providing insights from smart home power users, who are particularly attuned to building systems that offer flexibility and customization without sacrificing data privacy.

3. Method

While prior research has evaluated users' privacy needs in smart homes, this paper considers a distinct customer segment: privacy-conscious power users (PCPUs). We argue that these users' perspectives can be especially useful when considering how to better design SHDs and features to achieve the privacy needs described above. Power users are enthusiastic about devices and are willing to put in the time and effort to find solutions to mitigate their privacy concerns [33,57]. PCPUs likely have more experience trying out a range of devices and solutions to minimize data collection and maximize their ability to monitor and control data flows. They may also have insights into how these devices can better serve non-technical users, as they may have experience customizing their homes to accommodate non-primary users and bystanders.

This paper presents data from 10 focus groups with 32 privacy-conscious smart home power users. Focus groups are especially useful for developing a deeper understanding of how people with a shared experience feel about an issue [24]. Additionally, they enable a variety of perspectives and immediate follow-up from other participants and facilitators, which can be helpful for design-based inquiries and idea generation [24]. In our case, we used focus groups to bring together smart home enthusiasts who engaged in various approaches to managing their devices to understand their perspectives on the drawbacks of existing devices and interfaces. The group discussions allowed us to also solicit input into how to improve current and future smart home technology to accommodate diverse needs and provide users with greater awareness and management of data flows.

3.1. Recruitment and Study Design

This paper is part of a larger research project evaluating the privacy concerns and practices of smart home users. In summer 2021, we recruited people to participate in virtual focus groups to discuss their use of smart home technologies. After receiving IRB approval from the University of Maryland, we began posting recruitment messages on social media, including Twitter and smart home-related subreddits and Facebook Groups, inviting people who want to "*talk about how they use devices in their homes, the types of data these devices collect and share, and how we can design tools to better visualize this data and provide consumers with more control over their data*." The message directed potential participants to a short survey that collected demographics, details about their home environment, general privacy attitudes, and SHDs used.

We received 441 responses over one week; after removing spam responses, we had 277 potential focus group participants. We used two types of purposeful sampling—criterion and maximum variance [42]—to create a prioritized participant pool based on three factors. First, we looked at the devices respondents said they used. Given our interest in more advanced users, we removed from consideration anyone who selected a single type of device and prioritized those who used several different types of devices. Second, we looked at various items in the survey that would suggest a person was privacy conscious. This included attitudes toward privacy as well as managing devices to address privacy concerns. We prioritized people who reported engaging in privacy-enhancing device management (e.g., moving

Table 1: Participant IDs, Descriptive Data, and Smart Home Details

ID	Gender	Race	Age	Devices Used[1]	Integration Platform[2]	Advanced Network Management?[3]
P1	M	White	37	1,5,6,8,11	HK	No
P2	M	Black	37	1,2,3,5,6,7,8,11	HA, HK	Yes
P3	F	Black	45	1,2,4,5,6,8,10,11,12	HK	No
P4	M	White	48	1,8	none	Yes
P5	M	White	52	1,2,3,4,5,6,7,8,10,11	HK	No
P6	M	White	52	1,6,11,12	ST	No
P7	F	White	37	1,2,3,6,11	HK	No
P8	M	White	38	1,2,4,5,6,10,11	none	No
P9	M	White	35	1,2,3,4,5,6,7,8,10,11	HK	Yes
P10	n/a	n/a	39	1,2,3,4,5,6,7,10,11	HK	No
P11	M	White	34	2,3,4,6,7,11,12	HA	Yes
P12	M	White	27	5,6,8	HK	Yes
P13	M	East Asian	55	1,2,5,6,10,11	HK+HB	No
P14	M	White	40	1,6,8,11	HA	Yes
P15	M	White	33	1,3,4,5,6,7,8,10,11,12	HA	Yes
P16	M	E Asian, White	38	1,2,3,4,5,6,8,10,11,12	HK+HB	Yes
P17	M	White	24	1,2,6,8,11	none	No
P18	M	White	47	1,5,6,10,11,12	ST	Yes
P19	M	White	20	1,5,6,8,11,12	HK	Yes
P20	M	White	39	1,2,3,4,5,6,8,9,10,11,12	ST, HA	Yes
P21	F	White	29	1,2,6,8,10,11,12	HK, HA	No
P22	M	White	32	1,4,6,8,10,11,12	HK, HA	No
P23	NB	East Asian	25	1,6	none	No
P24	M	White	32	1,3,5,6,7,8,10,11	HK+HB	Yes
P25	M	White	40	1,2,4,6,8,10,11	HK+HB	Yes
P26	M	White	28	1,2,4,5,6,8,10,11,12	HK	No
P27	M	American Indian	36	1,2,5,6,8,10,11,12	HK	Yes
P28	M	E&S Asian	20	1,3,4,6,8	HK	No
P29	F	White	30	1,2,3,6,7,8,10,11	none	Yes
P30	F	White	42	1,2,3,4,6,10,11	HA, HK	Yes
P31	M	White	23	1,2,6,8	HK+HB	No
P32	M	E Asian, White	47	1,4,5,6,8,11	HK+HB	No

[1] Smart devices participants used: 1) speaker; 2) thermostat; 3) vacuum; 4) doorbell; 5) security camera; 6) lighting; 7) blinds; 8) TV 9) refrigerator; 10) door locks; 11) sensors; 12) other.
[2] Smart home hubs participants used: HK (Apple HomeKit), HB (Homebridge), HA (Home Assistant), ST (Samsung SmartThings).
[3] "Yes" is assigned to participants who used 1+ advanced network management strategies (i.e., setting up a Pi-hole or private DNS, flashing devices with custom firmware to run locally, setting up multiple routers to isolate devices, setting up firewalls).

devices out of private spaces, not using certain brands). Based on this, we had an initial set of 129 people we wanted to contact. We then applied the third criterion, which was to maximize diversity across gender, race, and home environment (e.g., own vs. rent; live alone vs. with others). This led to us prioritizing non-male and non-white respondents, who were under-represented in the pool.

Using our prioritized list, we began inviting people in batches to participate in 60-minute Zoom-based virtual focus groups in August 2021. We kept sessions small (3-4 participants) to ensure everyone had ample time to speak and to address the challenge of deciphering nonverbal communication virtually [50]. At least two authors attended each session. The research team debriefed after sessions and discussed if we were hearing new ideas and determining when we had reached data saturation [45]. In total, 82 people were contacted, 38 people signed up for a focus group, and 32 people participated in one of 10 sessions (see Table 1). Participants were compensated with a US$30 gift card.

Each session started with participants sharing general thoughts about their devices, how they built out their home environment, challenges and drawbacks they experienced, and concerns they had about devices. Participants then completed two brainstorming activities using Google Jamboard. We first asked them to map the types of data their devices collected onto a grid that captured the perceived sensitivity of that data along one axis and their desire to control and see data flows on the other. Following a discussion, we then asked them to brainstorm ideas and add post-it notes regarding the features they thought would be useful in visualizing or sharing data from their SHDs (see Figure 1; in this session, a team member organized post-its into clusters as participants added them). All participants described their Jamboards and the transcripts were analyzed. Due to time constrains, some sessions skipped the post-it note portion of the brainstorm session, moving straight to discussion. See appendix for full protocol.

3.2. Data Analysis

Audio from sessions was transcribed via Rev, then uploaded to Atlas.ti for qualitative coding. Using Miles, Huberman, and Saldaña's [38] approach to guide our analysis, we conducted two cycles of coding. We first developed an initial codebook based on the focus group protocol, the detailed notes taken during each session and our research questions. Each team member coded one transcript using the initial codebook, adding memos with questions and suggestions for new or collapsed codes. The team then discussed this initial process and refined the codebook. Following this, each transcript was then coded by two authors to ensure all relevant codes were applied.

Coded excerpts were then exported to Excel for secondary coding, following Braun and Clarke's [8] approach for thematic analysis, as well as Saldaña's [44] technique for "themeing the data." For each code, one team member reviewed all coded excerpts, taking notes on emergent patterns in the data. Through multiple rounds of reading and taking notes, team members began categorizing excerpts from each code and extracting themes, then writing a detailed analytic memo to describe each theme and provide examples [38]. These memos were discussed by the full team before organizing them into findings.

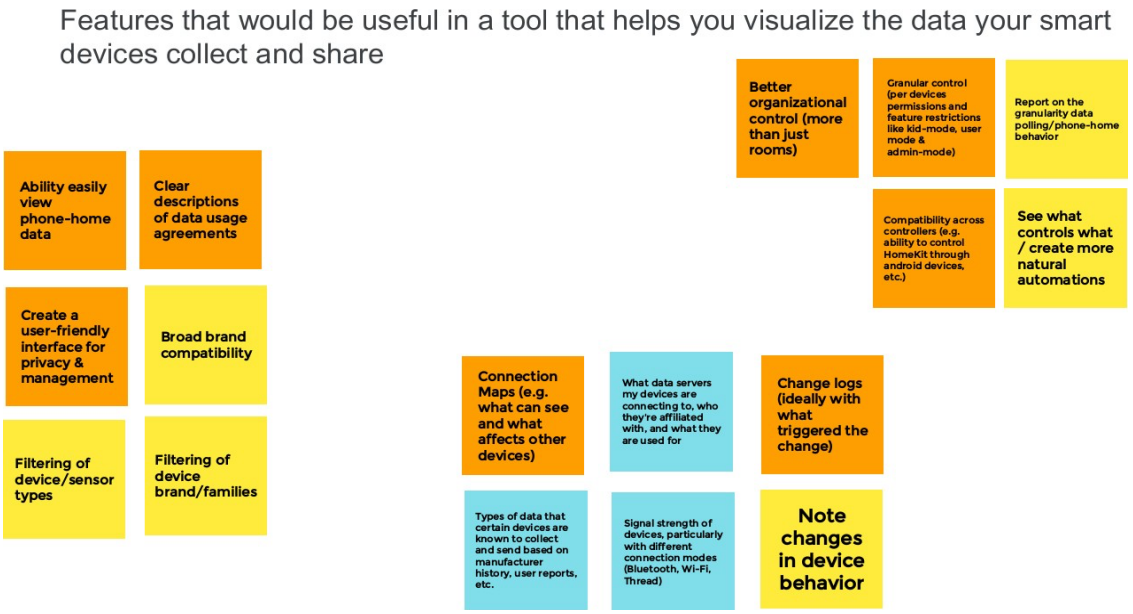


Figure 1. Jamboard screenshot from a focus group session, second design activity (brainstorming design features).

Table 2: Number of Participants Using SHDs/Hubs

Smart Home Device Usage			
Lighting	31	Speaker	30
Sensors	27	TV	23
Thermostat	19	Door locks	18
Security camera	18	Doorbell	15
Vacuum	13	Other	12
Blinds	8	Refrigerator	1
Smart Home Hub/Integration Platform Usage			
HomeKit	21	Home Assistant	8
Homebridge	6	Smart Things	3

In this paper, we focused primarily on two codes: Drawbacks (RQ1) and Design Features (RQ2). Additional codes were explored to supplement our analysis, including Bystanders and Data Concerns. We also conducted an additional round of analysis on the final brainstorming activity and Jamboards to further delineate the range of design features participants discussed into transparency, visibility, control and layers, taking note of why those features are important and the drawbacks they address.

4. Findings

4.1. Drawbacks and limitations to current smart home privacy and security management options (RQ1)

As our sample included many Apple HomeKit users, as well as participants who took complicated steps to manage data flows (see Tables 1 and 2), it is unsurprising they described conducting extensive research on privacy and security features before purchasing devices and using advanced configurations and customized settings during setup. That said, even with significant time and energy spent researching devices that aligned with their privacy and security needs, participants identified several drawbacks and limitations of smart home technologies currently available. Below, we describe three core tensions participants highlighted. Participants also raised general usability issues, but we focus solely on those connected to privacy and security.

4.1.1. Balancing privacy and security with functionality and interoperability

Most participants relied on HomeKit-compatible smart devices, with many noting they chose Apple because they trust the company and its commitment to data privacy and security. For example, P19 selected HomeKit because *"it's supposed to be really secure,"* and P13 indicated that when comparing smart device ecosystems, *"Apple was the one that had the most privacy built into it."* Participants also justified spending more money for HomeKit to achieve greater data security, with P5 saying: *"[Apple] costs you more, but you have that little bit more of a peace of mind that there's a little bit more control."*

Several participants specifically noted the privacy and security controls available for HomeKit-enabled routers, which offer a simple mobile interface that shows which devices are operating locally versus those connected to the internet. As P10 described, *"If you want it to have no internet access or just be able to connect to get firmware updates or whatever, there's different levels. It seems like the sort of thing that is ideal at the router level, because that's what's sitting between all your devices and the rest of the internet."* P19 also commented on this functionality, noting that the HomeKit control panel *"shows you all of your accessories, all of your hubs... Each accessory has the option to let it communicate freely, let it communicate to only a specific subset of domains that are strictly relevant to its operation, or only let it communicate locally."*

Others noted, however, that Apple's focus on simple interfaces can limit users' ability to manage data flows. P28 noted that default settings for devices may share more data than a user wants, while P19 acknowledged:

> *I feel like Apple largely has to appeal to the lowest common denominator… there are people who really want to get into the nuts and bolts of things, and that extra data is really valuable. And even though it might be a little overwhelming to the average user, it creates a value proposition there for people who are really, really interested in really protecting their privacy because that's their whole thing.*

Some participants acknowledged that enhanced privacy and security came with additional tradeoffs. Many HomeKit users who described using it for its enhanced privacy and security features also described frustration with its limited interoperability. P13 described challenges getting devices to work how he wanted them: *"To a certain extent, I'm sacrificing a lot of potential functionality and incurring greater cost, for the sake of being in a 'more private' environment."* This resonated with many participants, who noted general limitations when attempting to balance privacy and security with functionality and interoperability. P2 summarized this limitation, saying he spends significant time *"finding devices and figuring out the compatibility and the privacy"* and *"even then, I do buy some stuff that doesn't work the way I thought it would. That's really annoying."*

P2's comment reflects participants' experiences compromising their privacy and security preferences to achieve their automation needs. He said he bought Google Home Minis because he wanted to send text-to-speech commands through the HomePod, and the Minis had physical buttons for turning the microphone off. Likewise, P4 resorted to using an Amazon Echo for streaming music, saying, *"It wasn't my first choice for a smart device, but its capabilities*

were better than what I could find with my first choice [HomePod]. So yeah, that was a privacy tradeoff for me."

4.1.2. Balancing usability with customizability

Participants often sought to overcome limitations of interoperability by employing their technical skills, but HomeKit users faced additional drawbacks due to its general lack of customizability. Apple has long sought to simplify their products for greater usability, but for our PCPUs, this was seen as a drawback. P3 said she felt limited in what she could do in the Apple ecosystem: *"Being an old-school hacker geek, I don't like being told what I can't do. I like having the flexibility to play around."* P13 acknowledged that most people wanted (and needed) simple interfaces, but also wanted options for people with more advanced skill levels—who understood the risks and were willing to take them. He said, *"I wish Apple especially would be less about gatekeeping, and simplifying, and making things dumbed down for everybody at the expense of the few people who want to try more advanced things."* Likewise, P19 expressed frustration with the inability to easily tinker with devices, noting, *"If its biggest selling point wasn't that it was so secure, I would ditch it in a heartbeat."*

Participants described steps they took to overcome interoperability and customizability limitations with HomeKit by integrating third-party, open-source products like Homebridge or Pi-holes into their smart home ecosystem. These tools provide the ability to customize and control smart devices that may not be designed to work natively with the HomeKit ecosystem. They often require more effort and skill to install and configure than most non-power users have. P2 used a Pi-hole to block ads and track requests from external servers; one challenge he described was that *"you have to figure out which domain is associated with which device and how many there are."* P11 said one problem with these options is they're *"not necessarily user-friendly and you have to be willing to pick up coding or programming to make it work… I went into it without knowing how to program and I nearly threw a computer through a window at one point because of that."*

Participants often compared HomeKit to other products when describing customization limitations. P20, who used both Samsung's SmartThings as well as Home Assistant, said SmartThings *"had a great third-party dashboarding app I could build out an overview of my house and make it simple and easy for people to walk in and interact with."* Others suggested that attempts to balance usability with customizability have left everyone unsatisfied, such as P29: "*I think we're in this really weird state where, specific to smart home technology, it's a little too basic for the IT technology nerds, and a little too complex for the run of the mill user. So nobody's happy."*

4.1.3. More users, more challenges

A third set of drawbacks arose when multiple people were using devices. This complicated both smart device operation and participants' ability to manage privacy and security; in fact, many participants described struggling to balance their privacy and security needs with ensuring device usability and accessibility for other household members and guests.

The easiest solution for this challenge was to have a single household member (primary user) set up and manage devices. P19 said he and his partner came to an agreement where *"I'm just dealing with the whole thing. [My partner said,] 'I'm just trusting you with this. I'm not even going to try to understand. Just do it. I know you'll make it work.'"* Likewise, P30 said her husband trusts her with device setup and management because *"[he] knows that I'm a privacy person,"* while P26 said his wife lets him make purchasing decisions because she *"knows I care about the aesthetics of stuff, and I always run it by her."*

Multiple users can also create interoperability challenges. P25 noted their Apple HomeKit setup worked fine until a new (Android using) roommate joined the household: *"The thing about Apple devices in general is if you aren't in the Apple ecosystem, your friends with their 'dirty green bubbles,' they don't play well together. …I had to make a number of workarounds to make certain that everyone could still access devices."* P7 shared how she nearly switched ecosystems after purchasing a Sonos sound system, which wasn't compatible with HomeKit; in the end, she kept HomeKit rather than a more fragmented approach because she felt the latter would be less user-friendly for her daughter.

Having multiple users—including other household adults, children, and guests—led participants to consider ways to set up spaces with smart technology others could use while keeping them separate from the main ecosystem, especially when these non-primary users were not technically savvy. P22 used HomeKit as his primary hub, but he *"got an Alexa for just the guest room and got some lights that are connected over Bluetooth, so I feel confident they're local and offer that to the guest users,"* adding that this is because *"HomeKit does not have native support for limited guest users."* Multiple participants added physical (smart) switches as backups, especially for less-technical household members and guests. P16 struggled to find a smart switch for his partner that was not too complicated. He *"ended up picking a brand that acted like a light switch, and I did a lot of research to make sure that if it lost connection or if it failed to be smart, it could do everything dumb that it needed to do by itself."* P17 described *"exposing my partner to different things and easing them into that but having a physical fallback for them."* And P9 described negotiating with his less-technical wife: *"I think it's about creating off-ramps. When we moved into the new house, we just went with powered switches rather than smart bulbs. That way, at the end of the day, you can go over to the wall and hit a damn button and get the lights to turn on."*

4.1.4. Summarizing RQ1 findings

The drawbacks and limitations expressed by our power users reveal the challenges of balancing privacy and security with other features. Our participants were frustrated by difficulties in managing situations when multiple users might interact with smart devices. Some noted that if they are frustrated—with their knowledge, skills, and desire to tinker—then "average" users will be even more overwhelmed. P6 highlighted this sentiment: *"those who are not very tech-savvy, learning and understanding what does what, when you have lots of different devices, buttons, and lights and sensors, it can definitely overwhelm. Someone who's not technology-driven, you really need to create a cheat sheet for them."*

4.2. Design Feature Recommendations to Enhance Smart Home Ecosystems (RQ2)

RQ2 details the design features participants wanted to see developed or expanded that would address the drawbacks and limitations described in RQ1. In sharing their ideas, participants considered how such features would balance their needs and the needs of average users, which P29 captured when she said, *"I think there needs to be a better balance between dumbing things down so people can get [devices] up and running quickly, and being able to set up your smart home how you like it."*

Our participants wanted as much information as possible; however, they acknowledged that most users do not want or need so much information. Therein lays a challenge for designing tools that spanned all types of users; as P19 noted, *"There are people who really want to get into the nuts and bolts of things. And that extra data is really valuable. And even though it might be a little overwhelming to the average user, it creates a value proposition for people who are really interested in protecting their privacy because that's their whole thing."* At the same time, participants suggested that these features could help non-power users by *"bring[ing] the novice up to speed a little bit on what is really happening with their data"* (P6) and they would be *"something I would want to share with my family members or my partner, who I'm trying to convince"* (P17).

Below, we detail how participants balanced competing needs across different user types when describing features to enhance data transparency, visibility, and control.

4.2.1. Increase data transparency in a simple and standardized way to help users make informed decisions

Given that participants described investing significant time and energy researching SHDs, it is unsurprising they wanted device manufacturers and app developers to be more transparent regarding data collection and use practices to help them make more informed decisions both before purchasing and while using smart home technologies. Participants discussed two primary ways to make information more transparent: improved product labeling regarding data practices and improved notifications.

Our participants were largely dissatisfied with the limited information manufacturers and developers provided about data practices, and they wanted summary statements at multiple consumer touchpoints (device packaging, app, website). They believed that providing detailed information on data practices would help consumers make informed decisions *before* purchasing SHDs, including whether to purchase a device and where in their home they'd feel comfortable placing a device. P13 explained, *"You don't know until after you buy [a SHD] whether or not it's any good in terms of security or capability… so knowing ahead of time would be helpful."* Likewise, P31 said, *"the more information we can get as consumers, the better choices we can make.*

…give us all the information, be open, be transparent."

Participants described several core pieces of information that would aid them in decision making, including the types of data the device collects, how and when data is collected, and what purpose the data would be used for. Several participants specifically mentioned the need for privacy nutrition labels. P31 summarized the benefits of these labels, saying:

> *You look at your food, everything that's pre-packaged has nutritional labels on it, right? And I think if we could have some sort of electronic digital nutritional data like, "Hey, this uses this much electricity per hour, it sends out this kind of Z-Wave and Zigbee… this is the type of data we're collecting." I think that kind of nutritional label for electronics is what us, as consumers, are looking for.*

Participants liked that nutrition labels both provide key information for decision-making and do so in a clear, standardized way. P10 described Apple's privacy labels (which Apple unveiled eight months earlier) as *"distilling [information about data collection] down to something that's easily digestible and easy to compare one app to another";* they described wanting something similar for SHDs to make it *"easy to compare one device to another."* P17 reemphasized that the privacy labels should be concise and clear to help users parse out complex data practices. He said, *"I'd like a tool, like a nutrition label of sorts, that can say, 'Hey, this is the data that this device collects. This is why you should care or not care about it, and how often it does it.'"* The IoT privacy labels [16] reflect our participants' wants and needs for SHD data transparency. Additionally, participants suggested having an objective third party provide information about a device's data practices, rather than labels generated by companies, which resonates with the need for privacy ratings [21]. For example, P13 wanted *"a third-party reviewer or objective source to help folks who want to know more [information]."*

In addition to information about data practices, some partic- ipants wanted greater transparency regarding the identity of—and relationships between—device manufacturers and brands to show how data flows across other platforms or services shared by the same company. For example, P26 mentioned that *"some of these smaller companies like Aqara, a lot of stuff is just rebranded from a larger compa- ny from overseas that American consumers might not neces- sarily be aware of."* He said this was important information for making purchasing decisions because *"you might not want to use a specific company's products, but then you don't know that they've just white labeled that to somebody else."* P30 shared similar concerns, mentioning how the brand Tuya makes smart home technologies that *"all call home just by a random server in China"*—something she wanted to know before purchasing. Many participants de- scribed avoiding brands or devices from certain countries or companies because they didn't trust them to properly handle their data.

Participants also suggested ways to improve pop-ups and notifications when asking users to consent to data practices while setting up and managing SHDs. Notifications give users a chance to make informed decisions while using devices, but participants felt most lacked information to sufficiently inform and guide users. For example, P14 said, *"I find allow permissions like 'Do you allow permission for this app? That app?' woefully un-detailed. It's like, 'this app needs access to your camera.' Well, why does it need access to my camera? Why does it need access to my location? Some of the things just seem completely random."*

Like P14, many participants wanted better explanations for why devices were collecting data. P25 compared it to access requests on mobile apps, saying, *"it can give an explanation, ideally, like we need to access your water sensor to determine if the ground is wet outside."* Permissions could give users more information when deciding whether to agree to data collection. P32 summed up the idea by stating, *"if we could just have a little paragraph saying, 'this is what you're getting. This is where your data is going, and these are the people that are going to be using it,' then I can say, 'I'm cool with that' or 'I'm not cool with that.'"*

4.2.2. Greater visibility to manage device status and data flows

Along with increasing transparency about data collection and usage practices, participants underscored the importance of ongoing visibility into their smart home system and data flows, including device status; types of data being sent within and out of their network; where data is going; and how frequently data is being collected.

Building on their earlier comments regarding interoperability challenges (Section 4.1.1), participants expressed a strong desire for a standardized and centralized means to monitor their smart home data. P32 encapsulated this desire, saying: *"I want something that shows me everything in one place. Right now I have a smattering of ecosystems: Apple, Amazon, and Google. I would love for something in one app, something that I can just go and do everything."* To create a centralized tool for smart home data visibility, like P32 wanted, it would need to support a range of network protocols and technologies (e.g., HomeKit, Zigbee) to facilitate communication between devices. While Homebridge and Home Assistant already support this level of interoperability, our participants emphasized that these open-source tools require advanced networking skills beyond the out-of-the-box functionality associated with corporate ecosystems.

Within this centralized tool for managing data visibility, our participants wanted to quickly view current device status, see where they were located, and easily access their devices. Thinking about how to visually display device status, P15 noted a filtering mechanism: *"you might want to visualize it in a couple of ways. You might want to say, show me all the light switches in my house... or show me all the equipment that's operating in my kitchen right now, so it might not just be light switches in that case."* Participants also wanted to know how the house was automated and which devices or commands trigger other devices. Additionally, P28 wanted 'signal strength' to be in the overview as an indicator of device connection, to see *"if there are any particular problem spots where things might not be connecting properly."*

Several participants mentioned the importance of monitoring network traffic flow to know the data types and amount of data being shared by SHDs, as well as where device data went. P2 used a Pi-hole to *"track all the requests each site makes"* and wanted the feature to be *"as detailed as it could be and easily organized"* so he could see what data was being transmitted. P3 wanted *"to see what domains [device data] is going to and how much. I would love to see a time-of-day graph, where I could correlate, this is when I just got home so I'm seeing a large spike, and just the other day I wasn't even home and look what was going on."* Both participants' wants and needs build on to the idea of IoT Inspector [20], which visualizes network activities to identify security and privacy risks in the smart home environment.

As noted in Section 4.1, many participants wanted to keep as much of their data local as possible. In cases where a SHD required a cloud connection, participants like P10 wanted a *"visualization to show where in the world your data is going."* P10 further explained the broader concern motivating this feature request: *"I think some people might be surprised where some of their data is going, what countries it's going to. And you're like, why is my stuff going to that country? I just want to turn my lights on."*

To accompany visualizations of data flow such as where the data is being transmitted, our participants wanted additional information on *why* data was being collected. Thinking of features in a visualization tool for his parents, P9 focused on visualizing connections to services like Google AdWords. He said, *"the ability to visualize and understand [these con-*

nections], 'hey, you brought this thing online and it attempted to make 15 connections out. One was for a firmware update, one [was for] word upload logs, and the oth- ers were to shovel data out to known ad tracking agencies.'" P9 felt this information alone was simple and easy enough for his parents to interpret.

Like P9, many of our participants had other users in mind (e.g., bystanders) when thinking of ways to improve data visibility and control, which we will further elaborate in the next section. One reason our participants likely think of other users is they know firsthand the difficulty in parsing transaction logs and therefore the need for more user-friendly interfaces. As P24 noted: *"a lot of times I look at logs, or I look at lists of internet access, and I'm actually like, what is all this stuff, why are they going to all these different servers, and I just have no idea which server is which."* As such, even these advanced users don't always know why a device transmitted data to a certain domain.

4.2.3. Controls and notifications to manage smart home data

Building from the need for increased transparency and visibility, participants noted that these features were useful but still limited, especially when many devices and apps follow a "take it or leave it" approach—you can allow data collection, or you cannot use it. P22 brought this up when discussing privacy labels, saying, *"Apple's really good about saying, 'this app wants to access your photos' and allowing you control. But it's also like, well you have a nutrition label, but what's your options? I guess the question is: how low of a level can you disable stuff and still use the app? So I think that's part of the problem; it has a nutrition label and it's basically take it or leave it."* This dilemma between data minimization versus functionality is not new. For example, one of the central issues that Aretha [47] faced when providing a firewall as a control mechanism to manage data flows in smart homes was that it accidentally interfered with SHD operation. Additionally, our participants had many ideas for additional controls, including the ability to create allow/block lists for certain types of data to certain destinations, and more robust notification (e.g., alert, alarm) features that would allow users to manage data flows.

Many participants had advanced networking skills and used Pi-holes and private DNS setups to create lists of trusted and untrusted domains. That said, some expressed frustration with how these controls are often hindered by limited details on domain addresses, as well as the low accessibility of these tools. P2 captured this frustration when he said: *"If you try to wholesale block a device from connecting... it may stop working. But if you actually had the list of domains it was trying to hit, you could go through and block 90% of them and still keep your device working. So you're limiting your exposure while still getting the benefit of the device."* P15 described how greater network traffic visibility could lead to more impactful controls: *"Hopefully I'm able to say, look, this is only communicating with itself in my home hub. But maybe it's also communicating to my light switches for some reason—that's probably not a good idea. I should block that. Maybe it's communicating to some server that I don't know. I should block that."*

Others noted how better controls to create and manage block lists would provide peace of mind, such as P25's desire to *"easily put [IP addresses] in a box, if you don't feel so great about it"* or P24's wish that *"if in the certification process of the HomeKit app, or whatever they have to list to Apple, 'here's all the servers we're going to be needing to talk to and for this reason.' Then Apple says, 'okay, cool,' and they whitelist those and anything else gets blocked."*

Some participants wanted notifications (e.g., flags, alerts) to alarm users about unwanted activities based on the controls users created. For example, P25 suggested, *"if a Samsung device is contacting something other than a Samsung endpoint, that might be a red flag."* P4 suggested that after configuring a device, users could *"verify that with what the device is seeing on the network and say, 'Hey, you've told your TV, 'don't call home.' Hey look, it's calling home. Did you know that?'"* Once users are alerted to unusual activity on their network, participants wanted options to control their data, like P11, who mentioned *"a way to easily shut [data sharing] off, or shut off features, like, alright, here's the updates, but not sending out logs or…other stuff."*

Many participants wanted different settings for different types of users such as guests, domestic workers, and children. P22 mentioned that he used Alexas in guest rooms because Apple didn't yet have *"proper guest support."* P29 gave an example of babysitters (*"you want [babysitters] to be able to control lights or unlock the door, but you don't want them to be able to switch settings around")* and kids (*"you want [kids] to be able to control their room, but maybe not change the temperature level in the house").* On the other hand, P16, who self-described as a power user and said his wife had no interest in device management, said he'd rather take the initiative in a single location to manage data where *"I could go to and see a master status screen of everything going on in my house for the internet, the network traffic, data logs, power usage. I would love that."*

4.2.4. Summarizing RQ2 findings

While the PCPUs in our study benefited from their advanced technological skills to manage their smart device networks, they still looked to device manufacturers to provide better data transparency, visibility, and control. A common theme among our participants was that a user should not need special networking and security skills to understand how smart devices collect data or with whom data might be shared; manufacturers need to provide greater transparency and visibility as a default. Further, participants recognized that control over data flows they were able to create through customizations should similarly be available to all users regardless of technical proficiency. They

acknowledged that if they had to take extraordinary steps to see and manage data flows in their homes, then standard tools and interfaces will be insufficient to ensure usable data security and privacy controls for all.

5. Discussion

As the smart home landscape has expanded, usable privacy and security scholars have explored ways to make devices and automations more user friendly while encouraging expression of data rights like privacy, interoperability, and autonomy. In this study, we provide insights from privacy-conscious smart home power users who use both embedded features and third-party platforms to monitor traffic, build custom dashboards, and create custom domain block lists. By engaging with this unique set of users, we gained a deeper understanding of what features currently exist to help users understand and interact with their smart home data, the drawbacks of these tools and, most importantly, how the design of SHDs can be improved to increase the visibility and control of data flows to enhance user privacy—both for more technically savvy users as well as non-power users.

Our findings emphasize and extend many design features identified in prior work [6,10,11,17,19,21,47,53,54]. Our participants were uniquely well-positioned to identify critical limitations of existing privacy-preserving design features that prevented them from effectively mitigating privacy concerns and security risks raised by SHDs. Knowing that these power users—who were highly motivated to customize their homes to maximize benefits while minimizing external data flows—expressed significant frustration with currently available options signals an urgent need to develop more user-friendly features and controls that are accessible for everyday smart home users.

Below, we consider two aspects of smart home design that should be addressed to move beyond P29's sentiment that "nobody's happy" to one where all users can be both satisfied and confident regarding their ability to manage the privacy of their smart home data flows.

5.1. Designing for Context

A key takeaway from our findings is that PCPUs wanted smart home technologies that could better balance their functionality and interoperability needs with their desired level of privacy—and they wanted such features to be usable (and understandable) by non-power users too. In short, PCPUs recognized that transparency alone is insufficient, especially since it often takes certain technical skills to make sense of data flows within a smart home network. Rather, they pushed for a broader focus on providing all users with the ability to assess data flows in comprehensible ways and within certain contexts of use in their smart home environments.

This desire aligns well with the contextual approach to privacy championed by Nissenbaum [40,41], where the appropriateness of personal information flows is contextually bound by factors such as the actors and purposes for such flows. Our participants built custom features and setups to manage their privacy contextually—allowing some data to be collected and transmitted only within contexts deemed appropriate. A smart TV sending data to the manufacturer's IP address might be acceptable, but sharing the same data with an unknown actor was deemed inappropriate.

While transparency alone isn't sufficient, it remains important, and a major challenge when designing for context is the general lack of transparency from device providers regarding what data they collect and how they use it. As we note above, P14 bemoaned the lack of information from devices and apps regarding *why* they wanted certain permissions. This lack of transparency makes it more challenging to manage devices effectively. Many of the issues our participants raised, such as ensuring *all* users can easily understand what data is being collected, who and where it is being shared, and for what reasons, will require companies to provide contextual information in a structured data format.

However, things may be changing. New international standards like Matter [13] have the potential to increase device interoperability and ease the task of designing centralized data visualizations and controls that support all smart device manufactures—addressing some of the frustrations expressed by our participants. Beyond these standards, new regulatory measures may be required if companies still determine that such disclosures are not prudent based on existing market incentives. Recent proposals in the U.S. (e.g., American Data Privacy and Protection Act; Terms-of-service Labeling, Design and Readability Act) would require greater transparency and disclosures for how technology platforms collect and user data [15,37]. The future of such laws remains unclear, and we urge smart device companies to respond to the prompts of the PCPUs in this study in advance of any regulatory requirements.

5.2. Designing for Users at Different Skill Levels

A second design challenge speaks to a knowledge and skills barrier. Our participants' descriptions of how they managed their SHDs—often through advanced network management approaches or complicated automations—points to a need for simpler solutions that account for variations in contextual factors like who is interacting with devices (e.g., children, guests) and device location (e.g., a speaker in a bedroom is different than a speaker in the kitchen).

Our PCPUs repeatedly noted that any design enhancements that stem from their experiences and recommendations must be flexible for a diverse range of users and stakeholders. Their statements resonated with prior work suggesting that users want privacy tools that are simple, proactive, and provide more control options [21]. While our participants often wanted as much data as possible, they acknowledged that

most users would be overwhelmed with so much information—confirming experimental findings showing how new smart device users struggle with complex interfaces and dashboards [1]—and would benefit by simpler features to facilitate data control.

Specific to increasing data transparency and visibility, our participants recognized that for maximum usability, information about data practices should be provided in a summarized, digestible format, with the option for more information for those who seek it. Such a solution aligns with Emami-Naeini et al.'s [16] approach to *layered labels*, which provide two types of information: a primary layer containing the most important content, and a secondary layer containing more detailed information. Layered labels also have the potential to facilitate learning, encourage discussion of data flows with household members and bystanders, and prompt companies to be even more transparent.

Our participants further discussed expanding these labels to include more information about the company collecting data, including if they are part of a conglomerate, whether any ownership or branding changes have recently taken place, and what location data might be sent to. This aligns with prior work highlighting users' interest in the relationships between companies handling data [47]. All users would benefit from such expanded transparency within the smart home ecosystem; by taking a layered label approach, a range of information can be made available across multiple layers to avoid overwhelming users less interested in technical details.

In terms of smart home data visualization and control, our participants wanted a centralized location to monitor and control their smart home data, including device status and network traffic flow. Furthermore, participants wanted to create custom allow/deny network traffic lists and wanted notifications to be automated based on that list. Lastly, our participants wanted different modes of control for other users such as secondary users and bystanders (e.g., guests, visitors, children, and domestic workers).

Previous scholarship has explored smart home data visualization and control tools, and our participants' feedback offers insights for further development. IoT Inspector [20] labels smart home network traffic and produces tables and charts for users to monitor their smart home data. Our participants explained how this type of tool could be enhanced with filtering capabilities for device type, communication endpoints, etc. Similarly, Aretha [47] provides the daily ebbs and flows as well as aggregated smart home data to users; however, users found the control mechanism difficult to use because there were too many endpoints to comprehend. Our participants described struggling parsing out domain lists from smart home data management tools and suggested that manufacturers provide a default network traffic list that their products require to function. This will allow users of every technical skill level to start on and build upon creating their preferred network traffic list. Furthermore, to serve a variety of users and their various privacy preferences, assigning different roles and responsibilities might be an idea. One example of this is Kratos+ [48], a multi-user access control mechanism with a priority-based access-policy negotiation technique. Kratos+ applies a policy negotiation algorithm that automatically solves and optimizes conflicting user access requests based on users' set priorities on different devices. In addition to conflict resolution, users are notified when changes are made or when their requests are rejected. Although Kratos+ is de- signed for access controls, we can think of a similar mecha- nism to resolve conflicting privacy needs in smart homes.

5.3. Limitations

Participants were recruited largely through popular online discussion forums on Reddit and Facebook. This recruitment method increased the possibility of biases within our sample based on the socio-demographic characteristics of who are active in such online spaces. Future work could seek to obtain a more diverse set of smart device power users as well as seek out bystander viewpoints to directly assess their privacy concerns and strategies regarding exposure to smart devices.

6. Conclusion

With the growing adoption of smart home technologies, companies have emphasized making their products simple and user-friendly, often to the detriment of providing users with full transparency, visibility, and control over the data these devices capture and share. Complementing and extending previous studies that explore how everyday users of smart devices think about and address data privacy concerns, this paper engages specifically with privacy-conscious power users (PCPUs) to gain a clearer understanding of the steps taken by those with advanced technical skills to manage their smart homes. We identify design recommendations inspired by these power users and prompt device manufacturers to consider how such enhanced levels of data visibility and control should not be restricted only to those with the skills to customize their smart environments. The data privacy and security afforded by these suggestions should benefit all users.

The smart device ecosystem continues to evolve, and the growing use of artificial intelligence to better learn and adapt to users' behavior and preferences [4,46] only increases the need for the expanded collection of user data by device companies. At the same time, new standards promise to make smart devices more ubiquitous and easier to use, likely yielding in fewer opportunities for users to have full visibility or control into how data is collected and used. While the PCPUs in our study might make do, they also acknowledged that "nobody's happy" when it takes extensive technical skills to maintain privacy, or more typical users are left without usable means to manage their privacy.

References

1. Jacob Abbott, Jayati Dev, Donginn Kim, Shakthidhar Gopavaram, Meera Iyer, Shivani Sadam, Shrirang Mare, Tatiana Ringenberg, Vafa Andalibi, and L. Jean Camp. 2022. Privacy Lessons Learnt from Deploying an IoT Ecosystem in the Home. In *Proceedings of the 2022 European Symposium on Usable Security* (EuroUSEC '22), 98–110. https://doi.org/10.1145/3549015.3554205
2. Noura Abdi, Xiao Zhan, Kopo M. Ramokapane, and Jose Such. 2021. Privacy Norms for Smart Home Personal Assistants. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3411764.3445122
3. Imtiaz Ahmad, Rosta Farzan, Apu Kapadia, and Adam J. Lee. 2020. Tangible Privacy: Towards User-Centric Sensor Designs for Bystander Privacy. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2: 1–28. https://doi.org/10.1145/3415187
4. Amos. 2022. Artificial Intelligence Is the Next Step for Smart Homes. *Unite.AI*. Retrieved February 14, 2023 from https://www.unite.ai/artificial-intelligence-is-the-next-step-for-smart-homes/
5. Noah Apthorpe, Yan Shvartzshnaider, Arunesh Mathur, Dillon Reisman, and Nick Feamster. 2018. Discovering Smart Home Internet of Things Privacy Norms Using Contextual Integrity. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 2: 1–23. https://doi.org/10.1145/3214262
6. Nata Barbosa, Zhouhao Zhang, and Yang Wang. 2020. Do Privacy and Security Matter to Everyone? Quantifying and Clustering User-Centric Considerations About Smart Home Device Adoption. *Usenix*. Retrieved from https://www.usenix.org/conference/soups2020/presentation/barbosa
7. Bruhadeshwar Bezawada, Maalvika Bachani, Jordan Peterson, Hossein Shirazi, Indrakshi Ray, and Indrajit Ray. 2018. IoTSense: Behavioral Fingerprinting of IoT Devices. Retrieved from http://arxiv.org/abs/1804.03852
8. Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2: 77–101. https://doi.org/10.1191/1478088706qp063oa
9. Joseph Bugeja, Andreas Jacobsson, and Paul Davidsson. 2016. On Privacy and Security Challenges in Smart Connected Homes. In *2016 European Intelligence and Security Informatics Conference)*, 172–175. https://doi.org/10.1109/EISIC.2016.044
10. George Chalhoub, Martin J Kraemer, Norbert Nthala, and Ivan Flechais. 2021. "It did not give me an option to decline": A Longitudinal Analysis of the User Experi- ence of Security and Privacy in Smart Home Products. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–16. https://doi.org/10.1145/3411764.3445691
11. Chola Chhetri and Vivian Genaro Motti. 2022. User-Centric Privacy Controls for Smart Homes. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2: 349:1-349:36. https://doi.org/10.1145/3555769
12. Eun Kyoung Choe, Sunny Consolvo, Jaeyeon Jung, Beverly Harrison, and Julie A. Kientz. 2011. Living in a glass house: a survey of private moments in the home. In *Proceedings of the 13th international conference on Ubiquitous computing - UbiComp '11*, 41. https://doi.org/10.1145/2030112.2030118
13. Connectivity Standards Alliance. 2022. Matter Arrives Bringing A More Interoperable, Simple And Secure Internet Of Things to Life. *CSA-IOT*. Retrieved from https://csa-iot.org/newsroom/matter- arrives/
14. Julia C. Dunbar, Emily Bascom, Ashley Boone, and Alexis Hiniker. 2021. Is Someone Listening?: Audio-Related Privacy Perceptions and Design Recommendations from Guardians, Pragmatists, and Cynics. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 3: 1–23. https://doi.org/10.1145/3478091
15. Gilad Edelman. 2022. Congress Might Actually Pass ADPPA, the American Data Privacy and Protection Act | WIRED. *Wired*. Retrieved February 11, 2023 from https://www.wired.com/story/american-data-privacy-protection-act-adppa/
16. Pardis Emami-Naeini, Yuvraj Agarwal, Lorrie Faith Cranor, and Hanan Hibshi. 2020. Ask the Experts: What Should Be on an IoT Privacy and Security Label? In *2020 IEEE Symposium on Security and Privacy (SP)*, 447–464. https://doi.org/10.1109/SP40000.2020.00043
17. Pardis Emami-Naeini, Sruti Bhagavatula, Hana Habib, Martin Degeling, Lujo Bauer, Lorrie Faith Cranor, and Norman Sadeh. 2017. Privacy Expectations and Preferences in an IoT World. *Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017)*: 399--412.
18. Margaret Hagan. 2016. User-centered privacy communication design. In *Proceedings of the Symposium on Usable Privacy and Security*, 22–24. Retrieved from https://www.usenix.org/conference/soups2016/workshop-program/wfpn/presentation/hagan
19. Julie M. Haney, Susanne M. Furman, and Yasemin Acar. 2020. Smart Home Security and Privacy Mitigations: Consumer Perceptions, Practices, and Challenges. *NIST*. Retrieved from https://www.nist.gov/publications/smart-home-security-and-privacy-mitigations-consumer-perceptions-practices-and
20. Danny Yuxing Huang, Noah Apthorpe, Frank Li, Gunes Acar, and Nick Feamster. 2020. IoT Inspector: Crowdsourcing Labeled Network Traffic from Smart Home Devices at Scale. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 2: 1–21. https://doi.org/10.1145/3397333

21. Haojian Jin, Boyuan Guo, Rituparna Roychoudhury, Yaxing Yao, Swarun Kumar, Yuvraj Agarwal, and Jason I. Hong. 2022. Exploring the Needs of Users for Supporting Privacy-Protective Behaviors in Smart Homes. In *CHI Conference on Human Factors in Computing Systems*, 1–19. https://doi.org/10.1145/3491102.3517602
22. Patrick Gage Kelley, Joanna Bresee, Lorrie Faith Cranor, and Robert W. Reeder. 2009. A "nutrition label" for privacy. In *Proceedings of the 5th Symposium on Usable Privacy and Security - SOUPS '09*, 1. https://doi.org/10.1145/1572532.1572538
23. Euiyoung Kim, JungKyoon Yoon, Jieun Kwon, Tiffany Liaw, and Alice M. Agogino. 2019. From Innocent Irene to Parental Patrick: Framing User Characteristics and Personas to Design for Cybersecurity. *Proceedings of the Design Society: International Conference on Engineering Design* 1, 1: 1773–1782. https://doi.org/10.1017/dsi.2019.183
24. Richard A. Krueger and Mary Anne Casey. 2014. *Focus Groups: A Practical Guide for Applied Research*. SAGE Publications, Inc, Los Angeles.
25. Albrecht Kurze, Andreas Bischof, Sören Totzauer, Michael Storz, Maximilian Eibl, Margot Brereton, and Arne Berger. 2020. Guess the Data: Data Work to Understand How People Make Sense of and Use Simple Sensor Data from Homes. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–12. https://doi.org/10.1145/3313831.3376273
26. Josephine Lau, Benjamin Zimmerman, and Florian Schaub. 2018. Alexa, Are You Listening?: Privacy Perceptions, Concerns and Privacy-seeking Behaviors with Smart Speakers. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW: 1–31. https://doi.org/10.1145/3274371
27. Scott Lederer, Jennifer Mankoff, and Anind K. Dey. 2003. Who wants to know what when? privacy preference determinants in ubiquitous computing. In *CHI '03 extended abstracts on Human factors in computing systems,* 724-725. https://doi.org/10.1145/765891.765952
28. Hosub Lee and Alfred Kobsa. 2016. Understanding user privacy in Internet of Things environments. In *2016 IEEE 3rd World Forum on Internet of Things (WF-IoT)*, 407–412. https://doi.org/10.1109/WF-IoT.2016.7845392
29. Christian Leichsenring, Jiajun Yang, Jan Hammerschmidt, and Thomas Hermann. 2016. Challenges for smart environments in bathroom contexts. In *Proceedings of the 1st Workshop on Embodied Interaction with Smart Environments* (EISE '16), 1–7. https://doi.org/10.1145/3008028.3008033
30. Tianshi Li, Kayla Reiman, Yuvraj Agarwal, Lorrie Faith Cranor, and Jason I. Hong. 2022. Understanding Challenges for Developers to Create Accurate Privacy Nutrition Labels. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (CHI '22), 1–24. https://doi.org/10.1145/3491102.3502012
31. Wenda Li, Tan Yigitcanlar, Isil Erol, and Aaron Liu. 2021. Motivations, barriers and risks of smart home adoption: From systematic literature review to conceptual framework. *Energy Research & Social Science* 80: 102211. https://doi.org/10.1016/j.erss.2021.102211
32. Yuting Liao, Jessica Vitak, Priya Kumar, Michael Zimmer, and Katherine Kritikos. 2019. Understanding the Role of Privacy and Trust in Intelligent Personal Assistant Adoption. In *Information in Contemporary Society* (Lecture Notes in Computer Science), 102–113. https://doi.org/10.1007/978-3-030-15742-5_9
33. S. Marathe, S. Sundar, M. Bijvank, H. C. V. Vugt, and J. Veldhuis. 2007. Who are these power users anyway? Building a psychological profile. Retrieved July 6, 2022 from https://www.semanticscholar.org/paper/Who-are-these-power-users-anyway-Building-a-profile-Marathe-Sundar/1455563bf9242612c36f08e5a295aa139b8a1f04
34. Karola Marky, Sarah Prange, Max Mühlhäuser, and Florian Alt. 2021. Roles Matter! Understanding Differences in the Privacy Mental Models of Smart Home Visitors and Residents. In *20th International Conference on Mobile and Ubiquitous Multimedia*, 108–122. https://doi.org/10.1145/3490632.3490664
35. Faith McCreary, Alexandra Zafiroglu, and Heather Patterson. 2016. The Contextual Complexity of Privacy in Smart Homes and Smart Buildings. In *HCI in Business, Government, and Organizations: Information Systems*, Fiona Fui-Hoon Nah and Chuan-Hoo Tan (eds.). Springer International Publishing, Cham, 67–78. https://doi.org/10.1007/978-3-319-39399-5_7
36. Markus Miettinen, Samuel Marchal, Ibbad Hafeez, N. Asokan, Ahmad-Reza Sadeghi, and Sasu Tarkoma. 2017. IoT SENTINEL: Automated Device-Type Identification for Security Enforcement in IoT. In *2017 IEEE 37th International Conference on Distributed Computing Systems,* 2177–2184. https://doi.org/10.1109/ICDCS.2017.283
37. Carrie Mihalcik. 2022. TLDR Act aims to make website terms of service easier to understand. *CNET*. Retrieved from https://www.cnet.com/news/politics/tldr-act-aims-to-make-website-terms-of-service-easier-to-understand/
38. Matthew B. Miles, A. Michael Huberman, and Johnny Saldaña. 2013. *Qualitative Data Analysis: A Methods Sourcebook*. SAGE Publications, Inc, Los Angeles, CA.
39. Lily Hay Newman. 2020. Apple's App "Privacy Labels" Are Here—and They're a Big Step Forward. *Wired*. Retrieved from https://www.wired.com/story/apple-app-privacy-labels/
40. Helen Nissenbaum. 2004. Privacy as contextual integrity. *Washington Law Review* 79: 119–157. https://digitalcommons.law.uw.edu/wlr/vol79/iss1/10/

41. Helen Nissenbaum. 2010. *Privacy in context: Technology, policy, and the integrity of social life*. Stanford Law Books, Stanford, Calif.
42. Michael Quinn Patton. 2014. *Qualitative Research & Evaluation Methods: Integrating Theory and Practice*. SAGE Publications.
43. Sarah Prange, Ahmed Shams, Robin Piening, Yomna Abdelrahman, and Florian Alt. 2021. PriView– Exploring Visualisations to Support Users' Privacy Awareness. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (CHI '21), 1–18. https://doi.org/10.1145/3411764.3445067
44. Johnny Saldana. 2021. *The Coding Manual for Qualitative Researchers*. SAGE Publications Ltd, Los Angeles.
45. Benjamin Saunders, Julius Sim, Tom Kingstone, Shula Baker, Jackie Waterfield, Bernadette Bartlam, Heather Burroughs, and Clare Jinks. 2018. Saturation in qualitative research: exploring its conceptualization and operationalization. *Quality & Quantity* 52, 4: 1893–1907. https://doi.org/10.1007/s11135-017-0574-8
46. Samad Sepasgozar, Reyhaneh Karimi, Leila Farahzadi, Farimah Moezzi, Sara Shirowzhan, Sanee M. Ebrahimzadeh, Felix Hui, and Lu Aye. 2020. A Systematic Content Review of Artificial Intelligence and the Internet of Things Applications in Smart Home. *Applied Sciences* 10, 9: 3074. https://doi.org/10.3390/app10093074
47. William Seymour, Martin J. Kraemer, Reuben Binns, and Max Van Kleek. 2020. Informing the Design of Privacy-Empowering Tools for the Connected Home. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3313831.3376264
48. Amit Kumar Sikder, Leonardo Babun, Z. Berkay Celik, Hidayet Aksu, Patrick McDaniel, Engin Kirda, and A. Selcuk Uluagac. 2022. Who's Controlling My Device? Multi-User Multi-Device-Aware Access Control System for Shared Smart Home Environment. *ACM Transactions on Internet of Things*, 1–39. https://doi.org/10.1145/3543513
49. Yunpeng Song, Yun Huang, Zhongmin Cai, and Jason I. Hong. 2020. I'm All Eyes and Ears: Exploring Effective Locators for Privacy Awareness in IoT Scenarios. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (CHI '20), 1–13. https://doi.org/10.1145/3313831.3376585
50. David W. Stewart & Prem Shamdasani. 2017. Online Focus Groups, *Journal of Advertising*, 46:1, 48-60, https://doi.org/10.1080/00913367.2016.1252288
51. Neilly Tan, Richmond Wong, Audrey Desjardins, Sean Munson, and James Pierce. 2022. Monitoring Pets, Deterring Intruders, and Casually Spying on Neighbors: Everyday Uses of Smart Home Cameras. In *CHI Conference on Human Factors in Computing Systems*, 1–25. https://doi.org/10.1145/3491102.3517617
52. Parth Kirankumar Thakkar, Shijing He, Shiyu Xu, Danny Yuxing Huang, and Yaxing Yao. 2022. "It would probably turn into a social faux-pas": Users' and Bystanders' Preferences of Privacy Awareness Mechanisms in Smart Homes. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (CHI '22), 1–13. https://doi.org/10.1145/3491102.3502137
53. Yaxing Yao, Justin Reed Basdeo, Oriana Rosata Mcdonough, and Yang Wang. 2019. Privacy Perceptions and Designs of Bystanders in Smart Homes. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW: 1–24. https://doi.org/10.1145/3359161
54. Eric Zeng, Shrirang Mare, and Franziska Roesner. 2017. End user security & privacy concerns with smart homes. In *Proceedings of the Thirteenth USENIX Conference on Usable Privacy and Security* (SOUPS '17), 65–80.
55. Eric Zeng and Franziska Roesner. 2019. Understanding and Improving Security and Privacy in {Multi-User} Smart Homes: A Design Exploration and {In-Home} User Study. 159–176. https://www.usenix.org/conference/usenixsecurity19/presentation/zeng
56. Serena Zheng, Noah Apthorpe, Marshini Chetty, and Nick Feamster. 2018. User Perceptions of Smart Home IoT Privacy. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW: 1–20. https://doi.org/10.1145/3274469
57. Bu Zhong. 2013. From smartphones to iPad: Power users' disposition toward mobile media devices. *Computers in Human Behavior* 29, 4: 1742–1748. https://doi.org/10.1016/j.chb.2013.02.016

Appendix: Virtual Focus Group Protocol

Thank you for joining us today. The format of this session is a focus group. If you've never done one of these before, I have a set of questions I'd like to open up to discussion, but there's no formal method for answering. I encourage everyone to share their thoughts. My role is merely to facilitate the conversation; you all will be guiding it.

We're here today to talk about smart home technologies. This includes a wide range of devices, from smart thermostats and smart speakers, to connected TVs, fridges, vacuums, doorbells, security systems, toys, and more. We want to build a tool that helps consumers understand the types of data that are collected and used by these smart technologies, so the main goal of today's session is to learn from you about what information you think is important and what factors would make a tool like this useful to you.

We'll be recording the session today, but I want to assure you that whatever is being shared in this room today stays with us, and anything we use from this conversation will not be connected to your real name. That said, please treat this session as confidential and do not share things we discussed today with others. This session is scheduled to last 60 minutes. Does anyone have questions before we start?

Great, so let's get started. As a warm-up, let's talk about what types of smart technologies you use in your home. Can we go around the group and each of you share your name and then walk us through a normal day and **talk about the various devices you interact with and how you might use them. What do you like the most about them? What do you dislike about them?**

[Discussion]

For those of you who share your home with other people, smart devices may pose an interesting challenge because it's hard to "opt out" of using them. So we're curious, how do you make decisions about buying and using smart devices?

- Prompt (if needed): Is there one person who is "in charge" of managing these devices?
- Prompt (if needed): Do you have discussions with other household members before buying or setting up a device?

[Discussion]

Because you're using a range of devices, we can talk about building up an ecosystem of smart technologies that talk to each other and potentially share data. Thinking about that, one thing we want to hear more about is how you decide whether to connect smart devices to each other.

- Prompt (if needed): Do any of you struggle with the technical aspects of setting up and using these devices? Can you share an example of how that affected your decision on how to use a device?
- Prompt (if needed): Are there any devices you don't want to connect? Why?

[Discussion]

We're also interested in hearing about any times you've maybe been concerned about data being collected or transmitted by your devices. Are you ever worried about data being collected by one of your devices, or things a smart device might have "overheard" or collected without you knowing?

[Discussion]

Okay, we're going to spend the rest of the hour doing some design thinking activities. We're interested in ways to better share smart device data with consumers, and we want to think creatively about what that could look like.

[Design Thinking: Part 1]

Next, we want you to brainstorm all the types of data smart devices might collect about you and how you prioritize control over this data. Each of you have been assigned a Jamboard (virtual whiteboard) page. For the next few minutes list each type of data you can think of that is sent or received by each of your smart home devices (one per sticky note).

Place each sticky note on the grid provided. The grid has two axes capturing how sensitive a piece of data is to you and how much you want to be able to monitor and control that piece of data. So along the horizontal axis, place data you consider to be most sensitive on the right and along the vertical axis, place the data you would like to have more visualization or control over in the top half.

[Answer questions, give them three minutes to do this, then summarize the themes briefly and ask if anyone has things to add.]

[Design Thinking: Part 2]

To wrap up, I'd like you to get your thoughts on what types of features you would want in a tool that helps you visualize the data your smart devices collect and share. I realize this is kind of an abstract question, there are no wrong answers. You'll have three minutes to jot down as many feature ideas as you can come up with, then we can talk them through. *[time permitting, use Jamboard; otherwise, have a group discussion]*

[Facilitator note: As all the sticky notes are posted we can start to look for trends/themes that emerge and group them, this often leads to a more fruitful discussion. If short on time, skip the sticky noting part and just ask them to discuss features as a group.]

Wrap-up: *thank everyone for attending and let them know about getting gift cards and that we'll share results once this is written up.*

Exploring the Usability, Security, and Privacy of Smart Locks from the Perspective of the End User

Hussein Hazazi
University of North Carolina at Charlotte
hhazazi@uncc.edu

Mohamed Shehab
University of North Carolina at Charlotte
mshehab@uncc.edu

Abstract

Smart home devices have recently become a sought-after commodity among homeowners worldwide. Among these, smart locks have experienced a marked surge in market share, largely due to their role as a primary safeguard for homes and personal possessions. Various studies have delved into users' apprehensions regarding the usability, security, and privacy aspects of smart homes. However, research specifically addressing these facets concerning smart locks has been limited. To bridge this research gap, we undertook a semi-structured interview study with 29 participants, each of whom had been using smart locks for a minimum period of two months. Our aim was to uncover insights regarding any possible usability, security, or privacy concerns related to smart locks, drawing from their firsthand experiences. Our findings were multifaceted, shedding light on mitigation strategies employed by users to tackle their security and privacy concerns. Moreover, we investigated the lack of concern exhibited by some participants regarding certain security or privacy risks associated with the use of smart locks, and delved into the reasons underpinning such indifference. In addition, we explored the apparent unconcern displayed by some participants towards specific security or privacy risks linked with the use of smart locks.

1 Introduction

Over the past two decades, the Internet of Things (IoT) has seen a significant uptick in the complexity and range of its applications. These applications span various sectors, from healthcare and smart manufacturing to smart home solutions that aim to enhance users' quality of life by affording them greater control over their home devices. One of the emerging technologies within this space is smart locks, which were introduced as an advanced alternative to traditional locks [17]. These devices offer a broader array of features beyond mere door locking and unlocking. In recent years, the smart lock market has expanded and grown more competitive, leading to an array of diverse designs and operational characteristics being introduced [3]. As per the Statista Research Department [15], the global smart lock market size, valued at approximately 0.42 billion dollars in 2016, is predicted to exceed four billion dollars by 2027. Considering the anticipated market size and the critical role smart locks play as a primary line of defense against potential intruders, it's crucial to evaluate their usability, privacy, and security from the perspective of current users. Understanding these user evaluations can highlight potential areas of improvement, informing future design and functionality enhancements for these devices.

Several studies, such as [11,29–31], have assessed concerns related to the usability, security, and privacy of smart homes, primarily from the user's standpoint. While other researchers [13,21,28] have examined the issues and possible mitigation strategies related to the privacy and security of smart locks from the systems perspective, little research has been done on smart locks' usability, privacy, and security from the end user's perspective, creating a gap in the research. To address this, our study was carried out to investigate user perceptions of privacy, security, and usability associated with smart locks. As part of this study, we investigated the following research questions:

- RQ1: What aspects of the smart lock's design and functionalities make it appealing to users from a usability standpoint?
- RQ2: What privacy and security concerns do end users have regarding smart locks?
- RQ3: How do end users deal with their privacy and se-

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.
USENIX Symposium on Usable Privacy and Security (SOUPS) 2023.
August 6–8, 2023, Anaheim, CA, USA

curity concerns?

- RQ4: What are the end user's perceptions regarding how the security and privacy of smart locks can be improved?

To help us answer these questions, we conducted semi-structured interviews with 29 smart lock users who had used their locks for at least 2 months before the interview and had used their smart locks to share access with other users. Our main goal was to better understand their concerns related to different aspects of smart locks as well as how they deal with those concerns. In general, our work makes the following contributions:

- Provide a thorough analysis of the usability, security, and privacy of smart locks from the perspective of the end user which gives us an understanding of how to improve each of the three aspects.
- Demonstrate that more work needs to be done to increase consumer awareness regarding security and privacy issues related to smart locks.
- Offer suggestions and recommendations for improving the security and privacy of smart locks based on our analysis of participant feedback.

2 Related Work

2.1 Smart Locks Security and Privacy

The comprehensive analysis of smart locks, with a particular focus on usability, privacy, and security from the user's perspective, remains largely unexplored. Despite this, there are multiple studies conducted by researchers, which delve into the examination of the overall security and privacy of the various smart lock models. For instance, Ye et al. [28] analyzed the security facets of the August smart lock, highlighting potential threats that could compromise the security and privacy of users. Their analysis reveals that these locks are vulnerable to several types of attacks, including Denial of Service (DoS), and loopholes that could allow attackers to access the owner's personal information, thereby risking their privacy. In a similar vein, Ho et al. [13] carry out a security and privacy analysis of five different commercially available smart locks to identify potential vulnerabilities and suggest effective defenses against these. Their findings indicate that some of these locks could fall victim to state consistency attacks, relay attacks, and unwarranted unlocking, among other problems. They also offer potential defensive strategies against such breaches. Several other studies [2, 5, 14, 20, 21, 26] aim to enhance the security and privacy of smart locks by proposing innovative frameworks, utilizing technologies such as blockchain [5], facial recognition [14, 20], and a combination of steganography and cryptography [2]. There is also an acknowledgment of the deficiencies in the access control management systems currently employed in commercial smart locks. These deficiencies could potentially jeopardize the security and privacy of these devices. Xin et al. [26] proposed replacing the prevalent role-based system with an attribute-based access control system, which could enhance the granularity of access control within smart locks and address issues like state consistency attacks, unauthorized unlocking, and cascading deletion of permissions.

2.2 Smart Home User Studies

As a member of the smart home device family, smart locks share several common attributes and functions with their counterparts. Most notably, these devices are typically managed through a dedicated companion app and maintain access logs. Previous studies have investigated various facets of smart home device usability, exploring topics like the motivation behind investing in such devices and the impact they have on enhancing domestic life quality. In [6], a significant number of participants expressed that the adoption of smart home devices elevated their sense of security and control within their homes. Participants also identified additional incentives for adopting these devices, such as the convenience they offer and the sense of staying abreast of technological advancements. Another study [4] proposed that smart home devices are generally expected to outperform their traditional counterparts in terms of functionality. However, the reliance solely on smartphone control and the absence of manual control options for some of the simplest yet most frequently used features was found to heighten user frustration [7]. This reflects the necessity for a balance between technological advancement and user-friendly design in the development of smart home devices.

Earlier studies have delved into the security and privacy apprehensions of end-users concerning smart home devices. For instance, Haney et al. [11], in their study involving interviews with 40 smart home device users, sought to understand any security or privacy worries these users may harbor and the strategies they adopt to alleviate these concerns. Their findings pointed out that the principal worry for users centered around devices equipped with audio and video features potentially being breached, a sentiment echoed by Zheng et al. [31] in their study. This suggests that users may express less concern over the security implications associated with other smart home devices lacking audio or video capabilities, such as smart locks. A number of studies, such as Haney et al. [9] and Tabassum et al. [22], report a seeming lack of concern among certain users regarding the security and privacy aspects of smart home devices. However, this apparent lack of concern doesn't necessarily denote lack of awareness. In fact, the studies indicate that this lack of concern often stems from a trust in the device manufacturer's ability to rectify any security issues, or a belief among users that they are unlikely to be targets for potential attackers [29]. Some users expressed

that their concern was confined only to smart home devices located in sensitive areas within their homes, suggesting a nuanced understanding of privacy and security concerns in different contexts [30].

In [23], Tabassum et al. conducted a user study with 39 participants (18 owners of smart locks and video doorbells and 21 non-owners) to explore the users' perceptions of the configurations and controls available in smart locks and video doorbells. Some participants reported concerns regarding unauthorized attempts to unlock the smart lock but they mostly turn the notifications on in order to be alerted to such attempts. Other participants were also concerned about hacking attempts which might allow adversaries to remotely unlock the door and provide physical access to the home. For most of the security concerns, some participants stated that their only way to cope with those concerns was to put trust in the manufacturers' security measures. However, unlike [23], our study puts more focus on examining smart locks users' level of concern regarding specific aspects of the security and privacy of the smart locks as well as investigating the usage behaviors of smart lock users. Zlatolas et al. also conducted a survey study with 306 participants in order to get an insight into their security perceptions of IoT devices within the smart home [18]. The findings of the study revealed a positive impact of device vulnerability awareness on the perception of security importance. Meaning that users who were more aware of the security vulnerabilities of smart home devices also believed in the importance of implementing mitigation strategies in order to protect their smart home devices against possible security threats and vulnerabilities.

Our study results mostly align with previous work while identifying additional privacy and security concerns and mitigation strategies specific to smart locks. Furthermore, our study investigates the usage behaviors of the smart lock's end users.

3 Methodology

We conducted a semi-structured interview study with smart lock users in order to gain a deeper understanding of smart lock users' opinions on different aspects of the lock based on their experience using the lock and to explore their ideas about how the lock can be improved in terms of usability, privacy, and security.

3.1 Participants

We sought participants who had used their smart locks for at least two months and shared electronic keys (digital keys) with others (family members, neighbors, parcel delivery, etc.). The participants were recruited through a mass email sent to the students and employees at the university as well as an advertisement post on the SmartHomes sub-reddit on the Reddit forums. Potential participants were asked to fill out a screening survey which contained questions such as what type of smart locks they have, for how much time have they been using them, and how many people do they share the locks with. Such questions allowed us to verify the participants' eligibility to take part in the study. A total of 29 participants were recruited. Among the participants, 10 were males and 19 were females, and all of them live in the United States. Most of them (n=16) were in the age group of 26-35 while 10 participants were in the age group of 18-25 and 3 participants were in the age group of 36-50. The majority of participants (n=24) stated that they had been using at least one smart lock for more than 4 months while 5 other participants had used their locks for 2-4 months.

3.2 Procedure

A researcher contacted participants who were selected for the study based on the screening survey to arrange a date and time for the interview. According to each participant's preference, all interviews were conducted virtually over Zoom, Google Meet, or Webex. Interviews lasted about 40 minutes on average and each participant was given a $10 Amazon gift card for participating in the study. The study was approved by the university's Institutional Review Board (Protocol #21-0295). Each interview was divided into two sections. The first part focuses on exploring the usability aspect of the smart lock while the second part focuses more on the privacy and security aspect of smart locks. Each part contained open ended questions as well as Likert scale questions. Participants were asked to explain their reasons for choosing a particular answer in order to better understand their perspective. Towards the end of the privacy and security section of the interview, we ask the participants to watch a YouTube video that was prepared and uploaded by one of the researchers which contains a demonstration of 2 types of state consistency attacks that some smart locks are susceptible to. Once the participant finishes watching the video, the researcher asks them some questions regarding the two issues illustrated in the video.

3.3 Data Analysis

Each interview conducted was audio-recorded and subsequently transcribed for analysis. Our data collection was bifurcated into qualitative and quantitative components. The qualitative data was processed using an inductive coding approach. This procedure was carried out independently by two researchers who then engaged in discussions to finalize the coded data, thereby resolving any potential disagreements. The final codebook consisted of 13 main codes and 53 subcodes. The complete codebook is added in Appendix A.3. Turning to the quantitative data, our main approach involved the use of descriptive statistics, given that the bulk of our interview questions were not formulated to test for statistical significance among variables. Nevertheless, for the few

questions that did require a test of statistical significance, we employed the non-parametric Wilcoxon Signed Ranks Test, considering the data didn't adhere to a normal distribution pattern.

4 Results

4.1 Usage Behaviors

The purpose of this section of the paper is to identify the popularly used smart lock features as well as understand end users' usage behaviors. Investigating these aspects of smart locks leads to a broader understanding of what aspects of the smart lock's design and functionalities make it appealing to end users from a usability standpoint (RQ1).

4.1.1 Adopting a Smart Lock

As an emerging technology, smart locks have their strengths and weaknesses in terms of privacy, security and usability, especially when compared to traditional locks that homeowners are already familiar with. In response to a question about whether participants hesitated before switching to a smart lock from a traditional lock, 12 participants said that they had some concerns initially and that it took them some time to become convinced that adopting a smart lock was the right choice. The two main reasons behind the hesitation were price and security. A smart lock can cost up to ten times as much as a traditional lock, which can be a big financial commitment. The security of smart locks was also a big concern among some participants who hesitated before adopting a smart lock.

Asked why they chose to switch from a traditional lock to a smart lock, the majority of participants (n=20) said it was because of how convenient using a smart lock is compared to using a traditional lock, whereas only 8 participants cited security as a reason for using one.

4.1.2 Automation

By using communication protocols such as Zigbee and Z-Wave, smart home devices can communicate with each other to automate tasks. In spite of this, only 4 out of 29 participants created automation scenarios that utilized smart locks. P2, for example, has an automation scenario set up so that when an authorized user unlocks the smart lock, the home security alarm is automatically disabled without having to manually disable it every time a resident enters the house. Many automation scenarios can be set up using the smart lock to increase the level of convenience and security of a house, but most participants were not aware of the possibility of creating automation scenarios that include the smart lock.

Reason for turning on notifications	Count
Get alerts when the deadbolt is jammed	10
Get alerts about who is accessing the house	8
Get security alerts	4
Get battery alerts	1

Table 1: Reasons for enabling smart lock notifications.

4.1.3 Features and Capabilities

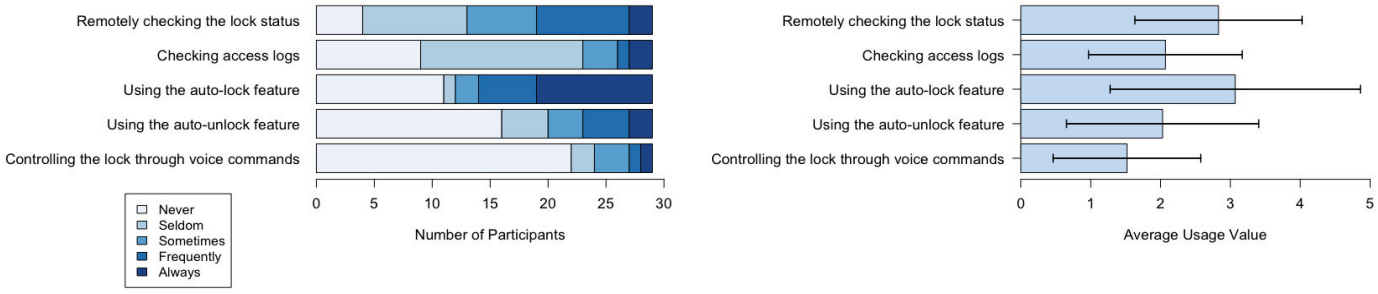
Compared to traditional locks, smart locks offer more features besides the basic function of locking and unlocking doors. The three most popular features that participants mentioned, unprompted, when asked to describe the features of their smart locks that they liked most were the ability to remotely control the lock (n=14), keyless entry (n=10), and the ease of giving others access (n=5). The ability to remotely check if the door is locked (n=3), the ability to unlock the door in multiple ways (n=3), and the auto lock feature (n=2) were not as popular among participants.

We also engaged the participants to evaluate their usage frequency of distinct smart lock features. To do this, we used a Likert scale that used the following designations: 'never', 'seldom', 'sometimes', 'frequently', and 'always', where 'never' corresponded to 1 and 'always' to 5. The "auto-lock" and the "remote lock status checking" features emerged as the most utilized among the smart lock features, as illustrated in Figure 1. The popular preference for these features stems from the heightened sense of security they afford to participants, particularly when they're away from home, by guaranteeing the door is securely locked – an observation underscored by a number of participants.

In terms of notifications, the majority of participants (n=21) stated that they keep smart lock notifications on. According to participants, the most common reason for enabling notifications is to be notified when the deadbolt jams on the door frame and does not lock properly, which is a common problem with smart locks. Notifications were also enabled to keep track of who was accessing the house in real-time, get alerts when the smart lock's battery was low, and see who was entering the apartment in real-time. in contrast, some participants (n=8) stated that they prefer to turn notifications off either because they don't prefer to use the app at all or because they find notifications annoying. Another participant was concerned that turning notifications on could violate other household members' privacy.

4.1.4 Managing Electronic Keys

Electronic keys are usually shared and revoked through the companion application. They can be in the form of a token on the user's smartphone or an access code that the user needs to enter every time they unlock the door. Participants were asked to evaluate two factors - ease of use and reliability -



(a) Usage frequency

(b) Average usage frequency

Figure 1: Smart lock's features usage frequency.

when it came to sharing their smart locks with others. Twenty-two participants found sharing access to the smart lock quite easy, but seven found it quite challenging, especially for older or less tech-savvy individuals. Among the participants, only two found it difficult to revoke someone's access to the smart lock. It is also worth mentioning that 13 participants reported that they never felt the need to revoke another person's access. Participants did not report any issues with the reliability of the access sharing process. When they share access with others, the other person is always able to operate the lock based on their access rights with no issues.

Access Sharing Patterns Access to the smart lock is usually shared through sending an invitation either by phone or email. When the other person accepts the invitation, they would be able to control the lock to the extent of their access level. Another way to share access to the smart lock is through an access code, usually 4 to 6 digits long, that allows the other person to unlock the door. Out of 29 participants, 13 reported that they only share access to their locks with people who live with them, such as roommates or family members. They feel more secure knowing that only the residents can unlock the door. The rest of the participants (n=16) stated that they give access to those who live inside the house, as well as others who don't live in the house such as guests, babysitters, contractors, dog walkers, etc. However, it is common for them to give "temporary access" to some of those who do not live in the house. For example, a dog walker who walks the dog from 10am to 11am can only unlock the door during these hours. Others, such as visiting family members or friends, can access the house at any time, but do not have full access to the lock in terms of checking access logs, giving access to others, or any other features besides locking/unlocking the door.

4.1.5 Usability Improvements

Although some participants were fairly satisfied with the smart lock's current features, others believed that it could be significantly improved by making some modifications and adding some new features. Some of these modifications include:

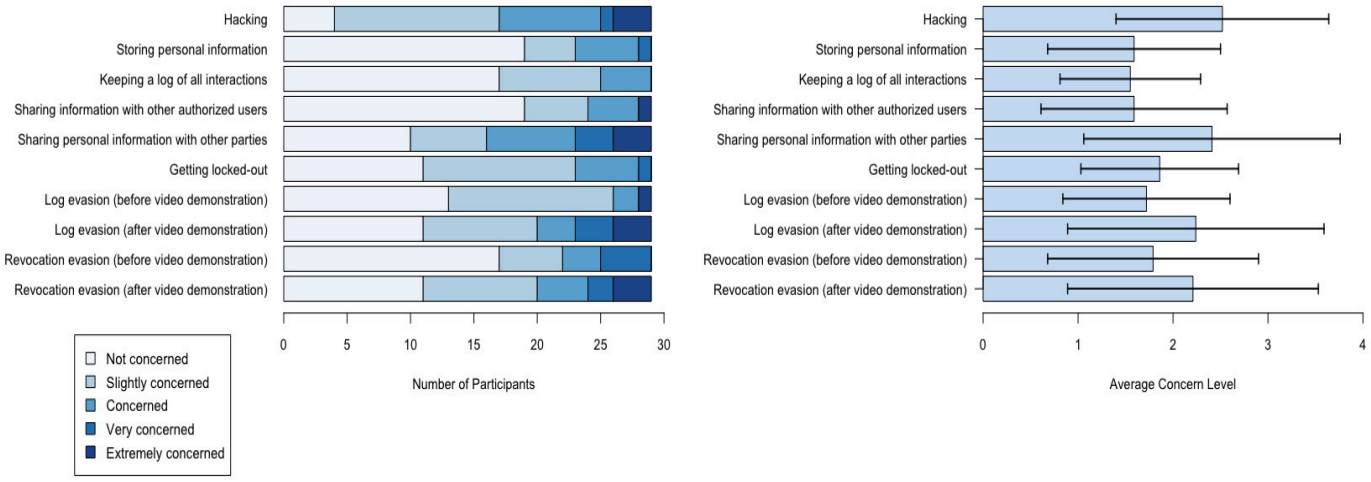
Improving the Battery In the case of smart locks, a dead battery can leave someone locked out of their home, especially if the lock doesn't offer any other means of unlocking it. Some participants (n=3) suggested different ways to improve the battery.

> **P27:** *"It would be nice if there was such thing as like a mini key fob that I could put on the bottom of the lock, just give it a charge so I can unlock it real quick to get into the house. That way, I could have that on my keys, and if I'm locked out when the battery's dead, I could just kind of like jump start it."*

Smart Watch Integration One of the participants suggested allowing smart locks to be operated by smart watches. This would be a very convenient feature especially for runners who prefer to leave their smartphones at home and only wear their smart watch. However, this feature already exists in some smart locks and watches such as the August smart locks that are compatible with Apple watches. Not all commercially available smart locks and smart watches support this feature though.

4.2 Security and Privacy Concerns

The purpose of this section is to explore and analyze the participants' insights regarding their privacy and security concerns (or lack thereof) with their smart locks (RQ2). In order



(a) Level of concern

(b) Average level of concern

Figure 2: The participants' level of concern associated with different smart locks privacy and security threats.

Security or privacy concern	Count
Hacking	9
Using and sharing access codes	7
Physical tampering with the lock	3
Losing the smartphone	2
Getting locked out	2
Revocation evasion	1

Table 2: The participants' security and privacy concerns related to smart locks (unprompted).

to ensure that they have had enough time to develop an opinion regarding the security and privacy controls of their smart locks, all participants have owned/used their smart locks for at least two months prior to the interview date and have used shared access to the lock with other users.

To gain an overall comprehension of the primary security and privacy concerns that smart lock users possess, we initially solicited from the participants any general security or privacy concerns they have associated with smart locks (Refer Table 2). This was followed by questions regarding their degree of concern about specific security and privacy issues related to smart locks (Refer Figure 2). The specific threats presented to the participants were formulated based on findings from previous research in the fields of smart locks and smart home security. These threats included concerns of log evasion, log revocation, and the possibility of being locked out, as discussed in previous studies such as [13, 16, 19, 25], which explored the security vulnerabilities prevalent in some smart lock systems. Furthermore, we asked the participants about concerns regarding hacking threats, storage of personal information, maintaining a log of all interactions, sharing personal details with other authorized users, and the possibility of information being disseminated to other parties. These additional concerns were also derived from previous research [11, 24, 29, 31], which delved into security concerns of smart home users associated with smart home devices. The participants' responses were recorded using a Likert scale, with designations ranging from 'not concerned' (assigned a score of 1) to 'extremely concerned' (score of 5).

4.2.1 Hacking

When asked, unprompted, about which privacy and security issue participants were concerned about the most when it comes to their smart locks, hacking was by far the most mentioned concern (N=9), which is in line with prior studies such as [11]. However, although 3 participants expressed extreme concern about hacking, a large portion of the participants were only slightly concerned (N=13) mostly because they don't believe themselves or their houses to be a potential or a high priority target for hackers, which seems to be a common thought process for a lot of homeowners [10, 22].

> **P22:** *"slightly concerned. I recognize that it can happen. But I don't see that our house is being a high priority target. It's not like we're particularly I don't feel like that we would be. I don't foresee us. Basically, security through obscurity is what I'm banking on. I don't see why anyone would want to get into our house specifically."*

4.2.2 Profiling and Information Collection

The majority of participants (n=19) expressed no concern about the smart lock collecting personal information about them and the residents of their home, which is consistent with previous research such as [11]. In fact, some participants appreciated that this sort of information is collected which can help improve the quality of the access logs. Some of those "not concerned" participants also believe that the information the lock collects is not significant and cannot harm them in any way, although when asked about the type of information they think the lock collects, some of them thought the lock only collects their name and email which is not accurate [1]. However, some other participants were more concerned about selling or sharing this information with other parties. P18, who was extremely concerned about sharing their information with third parties, says:

> **P18:** *"Sharing my privacy information with some other third parties is what I think is illegal and I don't feel it will be safe, because I trust that particular company and I don't trust the other."*

Based on the type of information the smart lock collects about its users and the fact that the smart lock also has the capability of sending and receiving information to and from other smart home devices, this creates the possibility of a profiling issue which is a huge privacy risk that most smart home users have to deal with. None of the participants explicitly mentioned "profiling" which could be because they are not familiar with that term or not even familiar with the type of information the smart lock collects and that it can lead to profiling. However, some participants were worried about others knowing the schedule of exactly when they are home and when they are not.

4.2.3 Using and Sharing Access Codes

Seven participants (n=7) expressed concern about the security implications of using or sharing their access codes. For example, two participants were concerned about an adversary observing them or other residents while using their access code to unlock their smart lock, which could allow the adversary to unlock their door later. Other participants (n=3) were more worried about the wear and tear of the keypads or touchpads that come with their smart locks (the most frequently used buttons wear faster than the others). Touchpads can show fingerprints, which can help an adversary figure out the access code based on observing how the keypad or touchpad looks based on which 4 buttons are used the most. Additionally, one participant was concerned about sharing access to the lock with others since they might not take security very seriously and make it easier for someone else to gain unauthorized access.

4.2.4 Physical Tampering with the Lock

The physical security of the smart lock was a concern for some participants (n=3). P5 is concerned about the lock itself being stolen for how expensive it is. Two other participants, on the other hand, were worried about the possibility of a burglar tampering with the lock and being able to gain access to the home. Especially in smart locks that have a physical keyhole as an extra option to unlock the door which can make it susceptible to picking just like traditional locks.

4.2.5 Losing the Smartphone

For a smart lock user, losing their smartphone is equivalent to losing their home key, especially if they do not secure their smartphone with a strong passcode or if they have the auto-unlock feature ON, which allows the lock to unlock itself when the smartphone is within a certain range of the lock without having to unlock the smartphone's passcode. Two participants were concerned that this could happen and an adversary could gain access to their homes. However, most smart locks already give their users the option to log in to their accounts through a website and disable the lost phone to avoid such an issue.

4.2.6 Getting Locked Out

Getting locked out of the home can be a huge security issue especially when it happens late at night or in a dangerous neighborhood. Although only about 28% of the participants (n=8) reported that they were locked out of their homes at least once because of the smart lock, the majority of participants (n=18) showed at least a slight concern that they might get locked out due to a smart lock related issue such as losing connectivity to the internet or a dead battery. Most of those who had already been locked out in the past also mentioned that it was indeed either an internet connectivity problem or a battery related issue.

4.2.7 Log Evasion and Revocation Evasion

Log evasion and revocation evasion affect smart locks that follow a Device-Gateway-Cloud architecture since they mostly rely on WiFi bridges or the user's smartphone to access the internet [13]. Through a companion app on the user's smartphone, these smart locks retrieve the access control list from a remote server, and verify it with the lock through Bluetooth to determine if a particular user is authorized to operate the lock. Unless the user's phone is connected to the internet or a WiFi bridge is available, the lock cannot retrieve the most recent access control list. As a result, even if user X's access to the smart lock was recently revoked, they can still operate the lock until the lock can connect to the internet and update the access control list. This is called revocation evasion which

Pre-Video (μ, σ)	Post-Video (μ, σ)	Z-value	P-value
Log Evasion Threat			
(1.72, 0.882)	(2.24, 1.354)	-2.334	0.020
Revocation Evasion			
(1.79, 1.114)	(2.21, 1.320)	-1.530	0.126

Table 3: The mean and standard deviation for the participants' level of concern regarding state consistency attacks in smart locks before and after watching the demonstration video.

is the first type of state consistency attacks. Likewise, a legitimate user, who has authorization to operate the lock, can also avoid appearing in the access logs simply by turning off their smartphone's internet connection. This is the second type of state consistency attack (evasion of access logs).

Although state consistency attacks have been heavily discussed in the literature [13, 16, 19, 25], only 3 participants stated that they were aware of the revocation evasion issue within smart locks while only 2 participants were aware of the log evasion issue. Users tend to be less concerned about security issues they are not familiar with. To give participants an overall understanding of the issues and how they can occur, we prepared a video demonstrating two types of state consistency attacks on one of the most popular smart locks on the market. We first asked the participants, on a scale of 1 to 5, how concerned they were regarding each of the two issues before watching the video and then again after watching the video towards the end of the interview. Our aim was to examine how raising the level of awareness of security threats affects users' level of concern about those threats.

The results showed an increase in the level of user concern regarding both of the security issues after watching the video as illustrated in Figure 1a and table 3. In order to determine whether statistically significant differences exist between the participants' level of concern before and after watching the video of the two security issues, a Wilcoxon Signed Ranks Test was performed. The tests revealed a statistically significant difference in the participants' level of concern in regards to log evasion (Z= -2.334, p=0.020, α= 0.05). However, The tests did not reveal a statistically significant difference in the participants' level of concern in regards to revocation evasion (Z= -1.530, p=0.126, α= 0.05). The reason behind this is that the participants were already more concerned about the possibility of revocation evasion compared to the possibility of log evasion even before knowing that the issues do exist. Therefore, although the participants' level of concern has mostly increased towards both issues after watching the video, it was more noticeable for log evasion.

After watching the video demonstration, most of the participants believed both issues to be very serious. However, they considered revocation evasion to be more serious compared to log evasion ($\bar{x} = 3.90$ and $\bar{x} = 3.49$, respectively). Referring to the revocation evasion problem, P13 says:

> **P13:** *"Extremely serious. That can really make or break someone's life extremely, especially with stalkers and domestic violence issues. I'm just trying to think about all the issues that someone has changed their locks because of some type of danger or harm that they felt that they might have been in to revoke someone's access into their home. So that person can still access their home, when they are not on Wi-Fi. That's scary."*

Furthermore, we asked the participants if they would switch back to traditional locks if they found that their smart locks had either of those problems. For both the revocation evasion and the log evasion issues, most participants (n=19 and n=23 respectively), stated that they would NOT go back to using a traditional lock. Some participants explained how they would buy a different smart lock instead of going back to a traditional lock because they appreciate the features that a smart lock offers. However, most of them stated that now that they know about those issues, they will make sure to test their smart locks and be more careful about which smart lock they buy in the future and who they share access to their locks with.

4.3 Reasons for the Lack of Concern

Using Mitigation Strategies Some participants mentioned that having added layers of security such as using a video doorbell or installing an alarm system on their smart locks was a factor that increased their trust in their smart locks and made them less concerned about possible security and privacy issues related to the smart locks.

Trusting Other Users Most of the participants who did not seem very concerned about most security issues related to smart locks stated that they only share access to their smart locks with people they absolutely trust and are not expecting any of these individuals to actively invade their privacy or compromise their security.

Trusting the Manufacturer The manufacturer's security and privacy policies play a crucial role in protecting the integrity and confidentiality of the data that is transferred from the end user to the manufacturer. Similar to previous research [22], Some participants stated that they trust the manufacturer to not sell or share their data with other third parties as well as keep their data secure on the cloud against any hacking attempts.

Everything about me is Already Out There! Some participants stated that their lack of concern with some privacy and security issues related to smart locks is due to the fact that their personal information is already on the internet one way or another and has already been sold to advertising agencies by other applications and services that they used in the past.

Therefore, they were not greatly concerned about their smart locks sharing personal information with other parties.

My House is not a Target! The participants were mostly aware of the fact that smart locks are susceptible to hacking. However, some of them did not show any concern regarding the possibility of hacking mainly because they were under the impression that hackers would have no interest in compromising their smart locks and gaining access to their homes.

4.4 Mitigation Strategies

Despite the fact that some participants showed concerns related to the security and privacy of smart locks, they also made it clear how convenient it is to use the smart lock and enjoy the added features compared to its counterpart the traditional lock especially when its counterpart also has its own security and privacy issues. However, the participants reported that they tend to use specific protective measures and mitigation strategies to cope with those concerns and improve the security of their smart locks without losing the convenience factor of using a smart lock (RQ3).

4.4.1 Adding Another Layer of Security

When asked if they use any other devices or gadgets to increase the security and privacy of their smart locks, most participants (n=25) stated that they do. The majority of those (n=24) have a video doorbell installed, which records everything that happens around the area where the smart lock is installed. In addition, it allows users to see who is actually at the door before unlocking it. The second most commonly used device to improve the security of smart locks was a chain guard or a swing guard (n=4), which is a small device that, when engaged, can be installed on the door and door frame to make it harder for an intruder to access the home even if they managed to get the smart lock to unlock. Two participants (n=2) also installed a secondary lock on the door, so that even if the smart lock was unlocked, the intruder would still have trouble getting in. Several participants (n=2) reported that their home had a security system that could alert them in case of a break-in. Those systems usually require the user to input a passcode every time they get through the front/back to stop the alarm from going off.

However, we asked the participants if they would still feel safe with the smart lock if those other security layers were not installed. To our surprise, 21 participants said they would, indicating either that they are confident in the security features of smart locks or that they do not consider their homes a target for intruders.

4.4.2 Configuring the Network

Some participants (n=3) suggested improving the security of the network that the smart lock connects to as a solution to concerns related to hacking and remote manipulation of the lock.

> **P12:** *"My biggest concern was the connectivity to the internet and, obviously, the ability that someone else may have to access the lock remotely, or gain access to the code or anything of that nature. I've kind of mitigated that by using Bluetooth instead of connecting it directly to wireless. And then when it's connected on my phone through Bluetooth, I actually have a separate wireless network that I'm connected to the separate VLAN so that anytime I'm connected to that device, it's not on the center VLAN that I use to surf the web and stuff like that."*

We hypothesized that improving authentication through using Multi-factor Authentication (MFA) would be something that at least some participants might mention as a possible mitigation strategy but when asked unprompted, none of the participants mentioned it. For this reason, we asked the participants if the applications they use to control their smart locks support MFA. About half of the participants (n=14) stated that their application does offer it, while 10 participants stated that they don't have this feature and 3 other participants did not know if they had it or not.

For the 14 participants who had access to MFA, 10 of them had it in the form of a One-Time-Password (OTP) that is sent to their phone or email when they log in from a new device, 5 participants have it in the form of a PIN, fingerprint, or face ID, that is required every time they use the companion application, and 1 participant had it in the form of a confirmation from an already logged in person. However, only one person out of the 14 participants who have the MFA feature stated that they use it frequently (in the form of a PIN, fingerprint, or face ID) while the others either don't use it or are required to use it every time they log in from a new device.

4.4.3 Managing Access Codes Carefully

Some participants (n=3) stated that they choose to manage access codes more carefully and put some regulations in place when it comes to creating and sharing access codes. This includes things like changing the access codes frequently and giving access only to a limited number of people who absolutely need it. Moreover, the companion applications used to control smart locks are usually reliable when it comes to sending out notifications of every interaction with the lock in real time to the homeowner as well as keeping an access log that records every interaction with the lock along with other information such as who interacted with the lock, when, and how. Some participants (n=2) said that this has been very

effective for them when it comes to dealing with their security concerns since they can always be notified of who is using the lock so they can confirm whether it was a person they recognize or not and can react to the situation accordingly.

4.4.4 Maintaining the Keypad/touchpad

As mentioned in the previous section, smart locks that are equipped with a touchpad/keypad have their own security issues especially when it comes to the wear and tear of the buttons and the touch screen itself. Participants (n=2) who have this sort of smart locks take some protective measures to deal with those possible security risks such as covering the touchpad/keypad with a plastic wrap so that it does not wear down as quickly as well as wiping off any fingerprints that it might catch after each use.

4.5 The Security of Smart Locks Compared to Traditional Locks

When asked whether it made them feel safer having a smart lock installed in their home compared to having a traditional lock, the majority of participants (n=19) said that it did. According to the participants, having features such as the ability to remotely lock the door, get security notifications, restrict others access time, and the ability to use the auto-lock feature made them feel that their home is secure even when they are away from home. However, other participants (n=10) did not necessarily feel more secure with the smart lock, but they appreciate its convenience. Some of them even felt less secure for various reasons such as the possibility of getting hacked, and the fact that others can see them as they type in their access codes and might be able to use that access code in the future.

4.6 Security and Privacy Improvements

In this section we report and discuss the participants' insights regarding how the smart lock's design and functionality can be altered in a way that enhances its overall security and privacy (RQ4).

4.6.1 Built-in Camera

Most commercially available smart locks don't have a built-in camera, but some of them can be easily integrated with other commercially available video doorbells. However, some participants (n=6) believe that having the doorbell camera already built-in can save the user money and time spent to integrate the two which sometimes might not even allow the user to use the full capabilities of both devices. Moreover, some participants lack the technological background to connect the two devices together. In fact, some participants (n=8) have both devices but do not have them connected due to different reasons such as not knowing how to connect them or the fact that they are not compatible in the first place. In terms of security, a built-in camera allows the users to see a video of who is interacting with the lock in real time as well as knowing exactly who is at the door before letting them in.

4.6.2 Improve Authentication

Some participants (n=6) believe that the authentication process within smart locks can be improved to increase the overall security of smart locks. According to the participants, they would feel more secure if instead of using an access code or a button on the companion application to authenticate, they would be able to use a more secure method such as face recognition or fingerprint (which is already available on some smart locks but not the most popular ones). However, some of participants also liked the idea of using Multi-factor Authentication (MFA) to improve the authentication process for logging into the companion application which was discussed at some point during the interview. Most of them were not familiar with the concept of MFA before the interview.

4.6.3 More Data Transparency

In line with previous work [27], several participants (n=5) believe that the manufacturer needs to be more transparent when it comes to explaining how the customer data is being used, who it's shared with, and how much of the user's information is shared.

> **P12:** *"I would say that it would be easier to have a little bit of better visibility into how your data is being used. It's not so transparent as to how your data is being used from third parties or from the company itself."*

4.6.4 Improving the Physical Security of the Lock

Two participants stated that the smart lock is not physically secure and could use some improvements in that aspect. This can be accomplished by implementing an intrusion detection system or a tamper detection system with specific sensors that can detect any tampering with the lock, attempts to break it, or hitting it with a strong force.

4.7 Limitations

Like many interview-based studies, our convenience sample size was limited and might not wholly reflect the broader population. Our recruitment efforts were predominantly focused on university students and employees, which confined us in terms of geographical diversity and the educational level of our participants. Therefore, nearly all our participants were from the United States with a generally high educational background. We attempted to address this lack of diversity by

promoting the study on Reddit forums. However, our attempt was hindered by the fact that most of the responses to the screening survey posted on Reddit came from bot accounts or were instances of a single person submitting multiple surveys. We identified this anomaly thanks to the data analysis and insights provided by the survey platform we utilized, Qualtrics.

5 Discussion

In this section, we will discuss some of the key takeaways from our study as well as discuss implications and recommendations for researchers and smart lock designers.

Smart Lock Adoption Our study revealed that most participants chose to adopt a smart lock mainly because the features that the smart lock offers make it more convenient compared to a traditional lock. This, however, contradicts with a prior study that aimed to explore the key factors affecting smart lock adoption in which improving the security and safety of the home was the most important factor that influenced the participants intention to adopt a smart lock [17]. This contradiction can be due to the different backgrounds or demographics of the participants in the two studies. Another reason could be the fact the participants in our study have had at least 2 months of experience using the smart lock before the interview, while the participants in the study conducted by Mamonov et al. hadn't adopted the smart lock at that point in time.

Convenience Over Security Although several participants expressed their concerns about privacy and security issues related to smart locks, most of them believed that the convenience of using the smart lock outweighs its security flaws. After all, its counterpart, the traditional lock, is not necessarily flawless in terms of security since it's susceptible to picking and tampering. However, several participants did not seem to be extremely concerned about the security drawbacks of the smart lock. Some of these participants were not aware of the possible security threats while others trust the mitigation strategies they put in place to increase the privacy and security of the lock and the smart home in general.

Unique Security Concerns and Mitigation Strategies Our findings revealed security concerns and mitigation strategies unique to smart locks which have not been discussed in prior studies that aimed to investigate the security and privacy concerns and mitigation strategies related to smart home devices in general. For example, some participants in our study expressed concerns regarding shoulder surfing attacks or the fact that attackers might be able to figure out the smart lock's correct access code based on which keys on the keypad are more worn due to being pressed more frequently. These sorts of concerns also introduced mitigation strategies that are more unique to smart locks such as maintaining the keypad/touchpad more regularly and managing access codes more carefully. Furthermore, some participants were also concerned about the possibility of losing their smartphone which would be equivalent to losing their key to the house, while other participants showed concerns regarding the possibility of getting locked out of their homes due to internet connection or battery related issues with the smart lock. While it's possible to mitigate some of the security concerns regarding most smart home devices by installing the device in a different location within the house, or turning the device off for a specific amount of time [22, 29], this is not applicable in the case of smart locks due to obvious reasons. However, our findings show that using an extra layer of security is the main mitigation strategy used by smart lock users to deal with their privacy and security concerns. For most participants, this extra layer of security was a video doorbell due to the fact that video doorbells are usually installed near the smart lock which provides the user with a clear view of what is happening around the lock and who is trying to interact with it.

The Trust Factor The lack of concern that some participants showed when answering questions related to security and privacy concerns was sometimes due to them having trust either in the other users, the manufacturer, or the security company that installed the smart lock [10, 29]. Having complete trust to the point of neglecting security vulnerabilities could be detrimental to the security of the entire home. For example, one participant mentioned that they do not check access logs because they trust all the other lock users. However, checking the access logs does not necessarily mean a lack of trust, but simply allows the lock owner to verify that only those who should have access to the lock actually do.

Sharing Electronic Keys The security of the smart lock and therefore that of the entire household, since compromising the smart lock can lead to unauthorized access to the home, is largely dependent on how safely the access codes and electronic keys are being managed. Carefully assigning access codes and electronic keys along with choosing the right access type for each person that uses the lock is extremely critical. For that reason, almost half of the participants chose to only give access to those who live in the house while the other participants, who gave access to non-residents, try to carefully choose the access level based on who needs access to the home, when, and why.

5.1 Implications and Recommendations

5.1.1 Design and Functionality Improvements

Access Control Management Currently, the majority of smart locks implement a Role-based Access Control (RBAC)

management system with 4 access levels: owner, resident, recurring guest, and temporary guest [13]. Each of the four access levels has specific access rights associated with it and the only two factors that the homeowner can manipulate when giving access to another user are the date and time (for the recurring guest and temporary guest access levels). However, more than half of the participants (n=16) stated that they share access to their smart locks with other users who don't live inside the house such as a babysitter, a pet walker, or a contractor. To improve the privacy and security of those who live inside the house, it's imperative to enable the homeowner to create more granular access control policies taking into account other environmental and contextual factors. For example, a homeowner might want the contractor to be able to use their access code only if no one is home to ensure the privacy of the home residents. Moreover, even when considering giving access to residents, prior studies, such as the study conducted by He et al., have proved that smart home users prefer to give access based on capability rather than device which also supports the need for more granular access control policies [12].

Video Doorbell Integration The fact that over 82% of the participants have a video doorbell installed next to their smart lock gives us an indication of how well these two devices complete each other and using them together can greatly improve the security and privacy of the household. However, many participants stated that although they have both devices, they don't necessarily have them connected either because they are not compatible, or because the user lacks the knowledge of how to connect them to get the most out of the two devices. We recommend, as well as many participants, that smart locks either have cameras already built-in or at least support seamless integration with other video doorbells in the market. The integration process needs to be simple with a clear and concise video tutorial to make it easier for those who are technically challenged to connect the two devices and get the added security and usability features.

Battery Many participants showed some concern regarding the battery life of the smart lock. Once the battery starts depleting, the lock becomes slower in responsiveness and sometimes does not even lock properly since it lacks the needed torque to properly lock the door. Prior work has indicated that smart locks suffer from sitting idle during extended periods of the day as well as having additional high peak current demands compared to other smart home devices [8]. Therefore, they require better power management in order to improve their battery life. Improving the battery life should be a priority along with increasing the frequency of battery level warnings that show on the user's smartphone before the lock gets to the stage where it struggles to unlock properly and not only when the battery is about to die completely.

5.1.2 Increasing Awareness

Our study shows that there is a general lack of awareness when it comes to security and privacy issues that the smart lock might be susceptible to. The lack of awareness often leads to lack of concern which can stop the smart lock user from implementing the correct protective measures and following the proper security practices to keep the lock secure. Therefore, more work needs to be done to educate the smart lock's user base about the possible security flaws and vulnerabilities. Our results show that the big majority of participants were not aware of state consistency attacks that some smart locks are susceptible to. Making them aware of those issues, however, has proved to increase the level of concern for some participants.

5.1.3 Transparency in Data Collection and Sharing

Our results revealed that the participants' level of concern regarding sharing their personal information with other parties is almost as high as the level of concern regarding hacking (Figure 2b). Therefore, it's imperative to give the users more control over what data is collected through the smart lock as well as more transparency about who gets access to such data. One way to improve the transparency in data collection and sharing is through adding more privacy controls and improving how privacy policies are displayed to the end user in a way that accommodates for users of different education levels, languages, and ages.

6 Conclusion

Given the continuous increase in the market size of smart locks year after year all over the world and the role smart locks play in maintaining the security and privacy of the household, more and more research needs to be done in order to improve the design and functionalities of smart locks. There have been numerous research papers published in the past discussing the security and privacy of smart locks from the perspective of the researchers, but little work has been done on the security and privacy of smart locks from the perspective of the end users. In this study, we focus on the end user's perspective of different aspects of the smart lock. We start our interviews by investigating the usage behaviors of smart locks' end users. We learned that big portion of smart lock users tend to share access to the lock with others who don't live in the house which justifies the need for improved access control policies. Our study also revealed that the convenience of smart locks was the number one factor in adopting a smart lock. The study also shows a lack of concern, as well as a lack of awareness, regarding some smart lock security and privacy threats.

Acknowledgments

We are extremely thankful to all the participants for their insights, time, and cooperation.

References

[1] August smart locks privacy policy | keeping your home & data locked down.

[2] Chaitanya Bapat, Ganesh Baleri, Shivani Inamdar, and Anant V Nimkar. Smart-lock security re-engineered using cryptography and steganography. In *International Symposium on Security in Computing and Communication*, pages 325–336. Springer, 2017.

[3] SANNE BJARTMAR HYLTA and PETRA SÖDERBERG. Smart locks for smart customers?: A study of the diffusion of smart locks in an urban area, 2017.

[4] Aykut Coskun, Gül Kaner, and İdil Bostan. Is smart home a necessity or a fantasy for the mainstream user? a study on users' expectations of smart household appliances. *International Journal of Design*, 12(1):7–20, 2018.

[5] Lucas de Camargo Silva, Mayra Samaniego, and Ralph Deters. Iot and blockchain for smart locks. In *2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, pages 0262–0269. IEEE, 2019.

[6] Luis Carlos Rubino de Oliveira, Andrew May, Val Mitchell, Mike Coleman, Tom Kane, and Steven Firth. Pre-installation challenges: classifying barriers to the introduction of smart home technology. In *EnviroInfo and ICT for Sustainability 2015*, pages 117–125. Atlantis Press, 2015.

[7] Christine Geeng and Franziska Roesner. Who's in control? interactions in multi-user smart homes. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2019.

[8] Chris Glaser and Aramis P Alvarez. Extending battery life in smart e-locks.

[9] Julie Haney, Susanne M Furman, Mary Theofanos, Yasemin Acar Fahl, et al. Perceptions of smart home privacy and security responsibility, concerns, and mitigations. 2019.

[10] Julie M Haney, Yasemin Acar, and Susanne Furman. " it's the company, the government, you and i": User perceptions of responsibility for smart home privacy and security. In *USENIX Security Symposium*, pages 411–428, 2021.

[11] Julie M Haney, Susanne M Furman, and Yasemin Acar. Smart home security and privacy mitigations: Consumer perceptions, practices, and challenges. In *International Conference on Human-Computer Interaction*, pages 393–411. Springer, 2020.

[12] Weijia He, Maximilian Golla, Roshni Padhi, Jordan Ofek, Markus Dürmuth, Earlence Fernandes, and Blase Ur. Rethinking access control and authentication for the home internet of things (iot). In *USENIX Security Symposium*, pages 255–272, 2018.

[13] Grant Ho, Derek Leung, Pratyush Mishra, Ashkan Hosseini, Dawn Song, and David Wagner. Smart locks: Lessons for securing commodity internet of things devices. In *Proceedings of the 11th ACM on Asia conference on computer and communications security*, pages 461–472, 2016.

[14] S Jahnavi and C Nandini. Smart anti-theft door locking system. In *2019 1st International Conference on Advanced Technologies in Intelligent Control, Environment, Computing & Communication Engineering (ICATIECE)*, pages 205–208. IEEE, 2019.

[15] Federica Laricchia. Global smart lock market size 2016-2027, Feb 2022.

[16] Yonglei Liu, Kun Hao, Jie Zhao, Li Wang, and Weilong Zhang. A novel smart lock protocol based on group signature. *International Journal of Network Security*, 24(1):130–139, 2022.

[17] Stanislav Mamonov and Raquel Benbunan-Fich. Unlocking the smart home: exploring key factors affecting the smart lock adoption intention. *Information Technology & People*, 34(2):835–861, 2020.

[18] Lili Nemec Zlatolas, Nataša Feher, and Marko Hölbl. Security perception of iot devices in smart homes. *Journal of Cybersecurity and Privacy*, 2(1):65–73, 2022.

[19] Saiprasanna Palle. *Smart Locks: Exploring Security Breaches and Access Extensions*. PhD thesis, Oklahoma State University, 2017.

[20] Varad Pandit, Prathamesh Majgaonkar, Pratik Meher, Shashank Sapaliga, and Sachin Bojewar. Intelligent security lock. In *2017 international conference on trends in electronics and informatics (ICEI)*, pages 713–716. IEEE, 2017.

[21] Bhagyesh Patil, Parjanya Vyas, and RK Shyamasundar. Secsmartlock: An architecture and protocol for designing secure smart locks. In *International Conference on Information Systems Security*, pages 24–43. Springer, 2018.

[22] Madiha Tabassum, Tomasz Kosinski, and Heather Richter Lipford. " i don't own the data": End user perceptions of smart home device data practices and risks. In *Fifteenth Symposium on Usable Privacy and Security (SOUPS 2019)*, pages 435–450, 2019.

[23] Madiha Tabassum and Heather Lipford. Exploring privacy implications of awareness and control mechanisms in smart home devices. *Proceedings on Privacy Enhancing Technologies*, 1:571–588, 2023.

[24] Blase Ur, Jaeyeon Jung, and Stuart Schechter. Intruders versus intrusiveness: teens' and parents' perspectives on home-entryway surveillance. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 129–139, 2014.

[25] Arvid Viderberg. Security evaluation of smart door locks, 2019.

[26] Zhenghao Xin, Liang Liu, and Gerhard Hancke. Aacs: Attribute-based access control mechanism for smart locks. *Symmetry*, 12(6):1050, 2020.

[27] Yaxing Yao, Justin Reed Basdeo, Smirity Kaushik, and Yang Wang. Defending my castle: A co-design study of privacy mechanisms for smart homes. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–12, 2019.

[28] Mengmei Ye, Nan Jiang, Hao Yang, and Qiben Yan. Security analysis of internet-of-things: A case study of august smart lock. In *2017 IEEE conference on computer communications workshops (INFOCOM WKSHPS)*, pages 499–504. IEEE, 2017.

[29] Eric Zeng, Shrirang Mare, and Franziska Roesner. End user security and privacy concerns with smart homes. In *Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017)*, pages 65–80, 2017.

[30] Eric Zeng and Franziska Roesner. Understanding and improving security and privacy in {Multi-User} smart homes: A design exploration and {In-Home} user study. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 159–176, 2019.

[31] Serena Zheng, Noah Apthorpe, Marshini Chetty, and Nick Feamster. User perceptions of smart home iot privacy. *Proceedings of the ACM on human-computer interaction*, 2(CSCW):1–20, 2018.

A APPENDICES

A.1 Screening Survey

- What is your first name?
- What is your email address?
- What age group do you belong to?
- What is your gender?
- What is your level of education?
- What is your current occupation?
- How many smart locks do you have installed where you live?
- Which smart lock(s) do you have installed where you live?
- Who installed the lock(s)?
- How long have you been using it (them)?
- How does your smart lock connect to the internet?
- How many people do you share access to the lock(s) with?
- Which virtual meeting platform do you prefer for conducting the interview?

A.2 Interview Questions

A.2.1 Smart Locks Usability

- What made you move from using a traditional lock to using a smart lock?
- Did you hesitate before making the move from using traditional locks to smart locks? Why?
- Do you have your smart lock connected to your video doorbell? Why?
- What would you say the top features of your smart lock that you mostly use?
- Who else can operate the smart lock, and what are their access levels?
- How easy do you find it to share keys with others? And how reliable?
- How easy do you find it to revoke other people's keys? And how reliable?
- Do you have notifications turned on for your smart lock app? Why?
- Do you connect your smart lock to other smart devices in your home using services like IFTTT? If yes, please talk more about the scenarios you have set up?
- Which aspects of the smart lock do you dislike or wish they would have been implemented differently?

- Compared to a traditional lock, how do you rate the locking/unlocking experience using a smart lock?
- In terms of locking/unlocking the door, how reliable is the smart lock compared to a traditional lock?
- How often do you find yourself checking the smart lock app on your phone to see if your door is locked/unlocked? (never, seldom, sometimes, frequently, always)
- How often do you find yourself checking the smart lock app on your phone to see the access logs? (never, seldom, sometimes, frequently, always)
- How often do you use your smart lock's auto lock feature? (never, seldom, sometimes, frequently, always)
- How often do you use your smart lock's auto unlock feature? (never, seldom, sometimes, frequently, always)
- How often do you control your smart lock using voice commands? (never, seldom, sometimes, frequently, always)
- Can you think of more features that you would like smart locks to have?

A.2.2 Privacy and Security Concerns Related to Smart Locks:

- What security or privacy related concerns do you have with your smart lock? How do you mitigate (deal with) those concerns?
- How concerned are you that your smart lock may malfunction and lock you out one day? (not concerned, slightly concerned, concerned, very concerned, extremely concerned)
- Has the smart lock ever locked you out of your home by accident? What was the reason?
- Would having a smart lock installed in your home make you feel safer compared to having a conventional lock? Why?
- How concerned are you that your smart lock might store your personal information and know your location at all times? (not concerned, slightly concerned, concerned, very concerned, extremely concerned)
- How concerned are you that the smart lock will keep a log of every time the lock is used along with the information of the person who used it? (not concerned, slightly concerned, concerned, very concerned, extremely concerned)
- How concerned are you that your smart lock might give others (such as your landlord) information about when you or your family members are home and when you are not? (not concerned, slightly concerned, concerned, very concerned, extremely concerned)
- How concerned are you that data collected by your smart lock might be shared with other parties? (not concerned, slightly concerned, concerned, very concerned, extremely concerned)
- How concerned are you that your smart lock might be hacked which allows unauthorized access to your home? (not concerned, slightly concerned, concerned, very concerned, extremely concerned)
- How concerned are you that the key revocation process might not be working correctly which allows others whose keys you have revoked to still have access to your home? (not concerned, slightly concerned, concerned, very concerned, extremely concerned)
- How concerned are you that some locking/unlocking activities might not appear on the smart lock's access logs? (not concerned, slightly concerned, concerned, very concerned, extremely concerned)
- What other security or privacy related concerns do you have with your smart lock?
- Do you use any other gadgets/ devices to increase the security of your smart lock?
- Does the app you use to control the smart lock allow you to use multi-factor authentication (MFA)? If yes, what form of MFA does the app offer? and how often do you use it?
- Do you think the companies that manufacture smart locks should add more features to make them more secure and increase the user's privacy? Could you give examples of such features?
- What other concerns do you have in regards to smart locks security and privacy?

A.2.3 Security Awareness

Each question listed below was asked twice, once for the revocation evasion security issue and another time for the log evasion security issue

- Did you already know that this issue existed?
- On a scale of 1 to 5, how serious do you think this issue is?
- On a scale of 1 to 5, how concerned are you that your smart lock might be affected by this issue?

- Would this issue cause you to go back to using a traditional lock instead of a smart lock?

A.3 Codebook

Code	Description
Motivation for adoption: Security	The core motivation to adopt a smart lock is to improve the security of the household
Motivation for adoption: Convenience	The core motivation to adopt a smart lock is to improve the convenience level within the home
Motivation for adoption: Did not personally install it	The user did not make the decision of purchasing and installing the smart lock (e.g., required by the landlord)
Motivation for adoption: Based on a recommendation	The user was motivated to adopt a smart lock based on a recommendation from other smart lock users
Hesitation to adopt: Price	The user hesitated before purchasing a smart lock because of its price
Hesitation to adopt: Security concerns	The user hesitated before purchasing a smart lock because of security or privacy concerns
Hesitation to adopt: Overwhelmed by the options	The user hesitated before purchasing a smart lock due to being overwhelmed by the different options on the market
Hesitation to adopt: Concerned about setup difficulties	The user hesitated before purchasing a smart lock because of concerns regarding the level of difficulty associated with its installation or setup
Most used features: Remote control	The user frequently locks and unlocks the door remotely
Most used features: Keyless entry	The user frequently unlocks the door without the need for a physical key
Most used features: Granting electronic keys	The user frequently grants access to other users electronically through the lock's companion application
Most used features: Remote status check	The user frequently checks the status of the lock to ensure that it's locked or unlocked
Most used features: Various unlocking options	The user unlocks the door through different methods such as using an access code, through the companion application, or using a fingerprint
Most used features: Auto-lock	The user configures the lock to automatically lock itself within a specific amount of time after being unlocked

Notifications on: Deadbolt Jammed	The user keeps the notifications on in order to get alerts when the deadbolt jams
Notifications on: Usage information	The user keeps the notifications on in order to get alerts about who is interacting with the smart lock, and when
Notifications on: Security alerts	The user keeps the notifications on in order to get security alerts such as the use of invalid access codes
Notifications on: Battery alerts	The user keeps the notifications on in order to get updates on the smart lock's battery level
Notifications off: Don't prefer using the companion application	The user does not get smart lock notifications because they don't prefer to use the companion application
Notifications off: Notifications can be annoying	The user turns the notifications off because they consider them to be annoying
Notifications off: Invasion of other residents' privacy	The user turns the notifications off to avoid invading the privacy of other home residents
Notifications off: Someone is always home	The user turns he notifications off because someone is always present at the house
Difficulty sharing access: Difficult for older or technologically challenged individuals	The user finds the process of sharing access to the smart lock to be difficult especially for older or technologically challenged individuals
Difficulty sharing access: All users must download the app	The user finds the process of sharing access to the smart lock to be difficult due to the fact that all the users need to download and configure the lock's companion application
Difficulty sharing access: Too many steps	The user finds the process of sharing access to the smart lock to be difficult due to the many steps the user needs to go through in order to grant the access
Access sharing patterns: Only share access with home residents	The user only shares access to the smart lock with those who live inside the house
Access sharing patterns: Share access with residents and non-residents	The user shares access to the smart lock with those who live inside the house as well as others who don't live inside the house
Usability improvements: Battery improvement	Improvements related to the smart lock's battery
Usability improvements: Smart watch integration	Allowing for a better integration between smart watches and smart locks
Privacy & security concerns: Hacking	Concerns about the possibility of hackers remotely manipulating the smart lock or gaining access to personal information
Privacy & security concerns: Shoulder surfing	Concerns about the possibility of others observing the user as he/she is using the access code to unlock the door
Privacy & security concerns: The wear and tear of keypads/touchpads	Concerns about the wear and tear of the most frequently used keys on the keypad/touchpad
Privacy & security concerns: Profiling	Concerns about other users or third parties obtaining information about who would be in the house (or not in the house) and when
Privacy & security concerns: Physical tampering with the lock	Concerns about physical tampering with the smart lock
Privacy & security concerns: Losing the smartphone	Concerns about a lost or stolen smartphone that can be used to operate the smart lock
Privacy & security concerns: Getting locked-out	Concerns about getting locked-out of the house
Privacy & security concerns: Revocation evasion	Concerns regarding the possibility that a revoked access might not be successfully revoked
Reasons for getting locked-out: Dead battery	The participant was locked out in the past due to a dead battery
Reasons for getting locked-out: Network or power issues	The participant was locked out in the past due to issues regarding the internet or a power outage

Reasons for getting locked-out: Auto-lock feature	The participant was locked out in the past due to the lock automatically locking itself while the phone is inside the house
Reasons for the lack of security concern: Using mitigation strategies	The user has showed a low level of concern about a possible security or privacy issue mainly due to them using a mitigation strategy to deal with possible threats
Reasons for the lack of security concern: Trusting other users	The user has showed a low level of concern about a possible security or privacy issue mainly due to having trust in other authorized users
Reasons for the lack of security concern: Trusting the manufacturer	The user has showed a low level of concern about a possible security or privacy issue mainly due to having trust in the security configurations the manufacturer has put in place
Reasons for the lack of security concern: Everything about me is already out there!	The user has showed a low level of concern about a possible security or privacy issue mainly because some of their personal information is already available to third parties
Reasons for the lack of security concern: My house is not a target!	The user has showed a low level of concern about a possible security or privacy issue mainly because they don't believe their house to be a target for hackers
Mitigation strategies: Adding another layer of security	The user installs other additional devices to increase the security of the home in case the smart lock is compromised
Mitigation strategies: Configuring the network	The user configures the network to improve the security of the smart lock
Mitigation strategies: Managing access codes carefully	The user creates and shares access codes carefully
Mitigation strategies: Maintaining the keypad/touchpad	The user maintains the keypad/touchpad so that it doesn't show more signs of wear and tear on the most frequently used keys
Security and privacy improvements: Built-in camera	Integrating a built-in camera can improve the security and privacy of the lock
Security and privacy improvements: Improve authentication	Implementing different authentication approaches can improve the security of the lock
Security and privacy improvements: More data transparency	More transparency about data collection and sharing would improve the users' privacy
Security and privacy improvements: Improving the physical security of the lock	Improving the physical security of the smart lock would improve the security of the overall security of the smart home

A.4 Demographics

Table A.1: Study participants demographic information

Participant	Gender	Age group	Education	Time spent using the smart lock	Connection to the internet
P1	Female	18-25	Bachelor's	More than 4 months	Directly (has a built in Wi-Fi)
P2	Male	26-35	Graduate student	More than 4 months	Wi-Fi hub (bridge)
P3	Female	26-35	Bachelor's	2-4 months	Wi-Fi hub (bridge)
P4	Male	26-35	Masters	More than 4 months	Not sure
P5	Male	26-35	Bachelor's	2-4 months	Wi-Fi hub (bridge)
P6	Female	18-25	-	More than 4 months	Directly (has a built in Wi-Fi)
P7	Male	26-35	Bachelor's	More than 4 months	Wi-Fi hub (bridge)
P8	Male	26-35	Bachelor's	More than 4 months	Directly (has a built in Wi-Fi)
P9	Female	26-35	Masters	More than 4 months	Directly (has a built in Wi-Fi)
P10	Female	26-35	Some college	More than 4 months	Wi-Fi hub (bridge)
P11	Female	36-50	Graduate degree	2-4 months	Not sure
P12	Male	26-35	Master's degree	More than 4 months	Smartphone's internet connection
P13	Female	26-35	Some college	More than 4 months	Not sure
P14	Female	26-35	Some college	More than 4 months	Not sure
P15	Female	18-25	Some college	2-4 months	Wi-Fi hub (bridge)
P16	Female	26-35	Some college	2-4 months	Directly (has a built in Wi-Fi)
P17	Female	18-25	Some college	More than 4 months	Not sure
P18	Female	18-25	Grad student	More than 4 months	Not sure
P19	Female	26-35	Grad student	More than 4 months	Not sure
P20	Male	36-50	Masters	More than 4 months	Wi-Fi hub (bridge)
P21	Female	18-25	Some college	More than 4 months	Directly (has a built in Wi-Fi)
P22	Male	26-35	Some college	More than 4 months	Not sure
P23	Male	36-50	PhD	2-4 months	Wi-Fi hub (bridge)
P24	Female	26-35	Master's	More than 4 months	Wi-Fi hub (bridge)
P25	Female	18-25	Some college	More than 4 months	Smartphone's internet connection
P26	Female	18-25	Some college	More than 4 months	Not sure
P27	Female	26-35	Associate degree	More than 4 months	Directly (has a built in Wi-Fi)
P28	Female	26-35	Grad student	More than 4 months	Wi-Fi hub (bridge)
P29	Male	18-25	Some college	More than 4 months	Directly (has a built in Wi-Fi)

"There will be less privacy, of course": How and why people in 10 countries expect AI will affect privacy in the future

Patrick Gage Kelley Celestina Cornejo* Lisa Hayes* Ellie Shuo Jin
Aaron Sedley Kurt Thomas Yongwei Yang Allison Woodruff
*Google, Ipsos**

Abstract

The public has many concerns and fears regarding artificial intelligence (AI). Some are general or existential, while others are more specific with personal repercussions, like weakened human relationships, job loss, and further erosion of privacy. In this work, we provide a deeper understanding of how AI privacy concerns are taking shape. We surveyed public opinion of AI's expected effects on privacy with 10,011 respondents spanning ten countries and six continents. We identify four main themes regarding how the public believes AI impacts privacy: vulnerability of data, highly personal data and inference, lack of consent, and surveillance and government use. Unlike many aspects of AI and algorithmic literacy, for which public perception is often reported to be riddled with inconsistency and misconceptions, these privacy concerns are well-reasoned and broadly aligned with expert narratives. Based on our findings, we provide a roadmap of public priorities to help guide researchers and the broader community in exploring solutions that ameliorate AI's impact on privacy, and to inform efforts related to civic participation.

1 Introduction

From facial recognition to smart home devices or self-driving cars, AI continues to spread quickly into people's daily lives. As people experience AI themselves, or hear about it in the media and through peers, they develop and refine their opinions of it. Researchers, corporations, governments, and public interest groups all seek to understand, measure, and potentially shape these opinions [14, 24–26, 51, 52, 73, 92, 103, 104]. Current assessments of public opinion of AI reveal both optimism about future benefits of AI as well as concerns about how AI may negatively affect people's lives and society in the future [5, 43, 63, 74], from questions about loss of human jobs to existential risks that AI may pose for humanity [5, 18, 38, 75, 96].

In this work, we focus on one particular concern that is commonly raised about AI: privacy [6, 56, 57, 82, 89]. Specifically, we explore how and why people believe AI will affect privacy in the future based on a survey of 10,011 respondents spanning ten countries and six continents (encompassing in total Australia, Brazil, China, Germany, Japan, Kenya, the Philippines, Russia, South Korea, and the United States). This contributes an international perspective on AI and privacy attitudes, including several countries in developing regions. We base our analysis on open-ended responses about how AI may affect privacy, supplemented by responses to closed-form questions. While the reader may expect there to be divergent views or misconceptions, many respondents expressed aspects of a coherent narrative that is broadly aligned with experts and privacy advocates. We found four main themes in all countries studied:

Data at Risk – Respondents believe that AI needs (lots of) data, which is gathered from multiple devices, crosslinked, aggregated, and made available online, where it is vulnerable to misuse and hackers.

Highly Personal – Respondents express that this large-scale collection includes highly personal data, which can be used to develop precise, personal insights that can be leveraged to influence or manipulate people for commercial or other purposes.

Without Consent – Respondents feel this scaled, personal data collection occurs without meaningful consent, and often without awareness, and they are often required to provide data to get access to useful AI services.

State and Surveillance – Our respondents identify ways that AI supports surveillance and governments through omnipresent monitoring and identification.

In light of these themes, we discuss how researchers and the broader community can work to mitigate the privacy risks of AI. Potential solutions range from technical design—such as the adoption of differential privacy or federated learning to minimize sensitive data—to privacy policies and platform adoption of AI principles. Critically, our findings show that the public has nuanced, well-reasoned concerns around pri-

vacy and AI that enable civic engagement and participatory democracy in shaping the future of privacy and AI.

2 Background

Much of the research on public perception of AI has been survey-based, often conducted in Western, English-speaking countries such as the US and the UK [14, 25, 38, 75, 104] but also in other regions or globally [5, 43, 63, 74, 92, 103]. Respondents typically expect AI will have a significant impact on the future, and often anticipate that its effects will be positive, with the most favorable impressions in emerging and/or Asian markets and more negative impressions (particularly recently) in the countries such as the US [5, 38, 43, 63, 74, 75, 92, 103]. At the same time, AI is neither interpreted as exclusively beneficial nor exclusively disadvantageous, and public response often indicates contradictory emotions [14, 56, 57, 73]. Privacy, job loss, increased social isolation, and other social topics have been highlighted as key concerns in surveys on AI [6, 38, 56, 57, 82, 89], and privacy has also been highlighted as a concern (either at a high level or in some cases specifically, e.g., government surveillance or lack of control) in surveys on autonomous vehicles, connectedness, facial recognition, IoT, personal data collection, and smart speakers [4, 8, 10, 13, 20, 51, 59, 65, 68, 102]. Some surveys have shown public support for responsible development and regulation of AI to address concerns [38, 92, 104].

Qualitative work has explored public perception of algorithmic systems, for example, finding that perception of algorithmic systems can vary substantially by individual factors or platform [37], and that end users often have fundamental questions or misconceptions about technical details of their operation [19, 39, 81, 90, 94, 95]. Qualitative studies with smart home device users primarily in the US and UK revealed privacy concerns such as constant monitoring, other parties' use of their data, or consent [1, 29, 49, 61, 71, 106]. These studies also reported that users had an incomplete or inadequate understanding of technical aspects of the systems' operation, particularly related to data processing, storage, and sharing.

AI is not only heavily discussed in academia, but is also a popular topic in public media and entertainment [27, 38], and studies have shown the public is likely to get information about AI from movies, TV, and social media [14, 28]. While researchers have argued that media narratives and fiction may be disproportionately frightening, especially in Western, English-speaking regions [26], studies have suggested that news reports may be more balanced or appropriately critical [32, 40, 78]. The popular press often features stories related to AI and privacy [3, 17, 30, 46, 48, 55, 64, 67, 70, 72, 84, 99], and privacy has been identified as a key concept in newspaper reports on AI [32]. Research has considered how media affects public opinion on privacy concerns such as government surveillance, data sharing, and companies' use of social media content [36, 41, 87], and smart home study participants have shared that their privacy concerns have been influenced by news reports and social media [29, 49].

Overall, our work sits within a growing body of research on people's perceptions of AI, across disciplines including critical studies, HCI, law, marketing, policy, psychology, usable privacy and security, and more. Perception of AI is highly complex, multi-dimensional, and far from fully understood. Methodologically, this means that techniques such as *triangulation* (studying the same phenomenon from multiple vantage points, in order to cross-check and more fully capture richness and complexity, e.g. using both qualitative and quantitative methods to see if the findings are consistent) [85] and *replication* (the reproduction and extension of prior work) [98] are particularly useful for this topic. Accordingly, we seek to broaden and enrich the understanding of people's perception of the relationship between AI and privacy by looking for emergent themes in a large number of open-ended responses from a wide range of countries.

3 Methodology

In order to better understand public perception of AI, we partnered with Ipsos, a global market research firm, to field our survey in August 2021. Methodologically, this work falls in the genre of public opinion polling, as described below. Our study plan was reviewed by experts at our institution in domains including ethics, human subjects research, policy, legal, and privacy. Our institution does not have an IRB, though we adhere to similarly strict standards.

3.1 Instrument Development and Translation

The survey instrument builds on previous versions which we deployed in 2018 and 2019 [56, 57]. To develop concepts and questions for all versions, we consulted experts at our institutions, reviewed published work, and drew on our own previous unpublished research. The 2021 version has some substantial modifications from previous versions, including the addition of an open-ended question about privacy which is the focus of this paper. Many questions in the final instrument were written uniquely for this survey while others were modified from or replicate other questions in the literature or the canon of public opinion surveys. In order to more accurately reflect real-world settings, we did not define AI, and left interpretation of the term to the respondents.[1] We did ask respondents two questions that serve as a knowledge check, which provide us some assessment of people's familiarity and understanding of AI. We included primarily closed-form questions as well as a few open-ended questions for free responses. We also included standard demographic questions such as age, gender, education, income, region, and urbanicity. The final

[1] In 2018, we had two versions of the survey (one that defined AI and one that did not) and responses to subsequent questions were similar regardless of whether a definition had been provided.

instrument included several dozen questions on topics related to artificial intelligence (see Appendix, Section 7).

After completing the instrument in English, we engaged cApStAn, a linguistic quality assurance agency with expertise in survey translation which had also partnered with us on the previous translations. cApStAn provided a translation style guide consistent with the previous rounds and identified complexities for particular concepts and languages. Ipsos' in-country translation teams and third party vendors referred to this guidance while translating the instrument to all target languages and iterated with cApStAn to finalize. Legacy translations were preserved when question/language pairs were identical to previous versions. See Appendix, Table 3 for languages offered. After fielding was complete, the responses were coded in-language as described below.

3.2 Deployment

We selected a range of countries with different characteristics, such as stage of technological development, nature of the workforce, and varied development indices. The survey was fielded to online panels (groups of respondents who have agreed to participate in surveys over a period of time) representative of the online population in each country. Consistent with the best panels available for online market research, such panels tend to be broadly representative of the general population in countries with high access to technology, but less representative of the general population in countries with more limited access to technology; for example, in developing countries they tend to skew urban. Respondents were recruited using stratified sampling (a method of recruiting specific numbers of participants within demographic subgroups), with hard quotas on age[2] and gender in each country.[3] The median survey length was 27 minutes across all completions. All respondents received incentives in a point system or cash at an industry-standard amount for their market. A summary of countries and demographics is provided in the Appendix, Table 3.

3.3 Data Processing and Analysis

Quality Checks. Ipsos conducted quantitative and qualitative checks to remove low quality responses on an ongoing basis until the quota was reached in each country. Example grounds for removal included being identified as a bot, speeding (answering substantially more quickly than the median time), or providing nonsensical or profane responses to open-ended questions. Overall Ipsos removed and replaced 9.4% of responses for quality.

[2] Ages ranged from 16 to 85, with a small recruit of 16 and 17 year olds in each country (between 19 to 80 youth participants per country), who participated with parental consent.

[3] The US was the only exception since the panel there operates by sending the survey to a representative sample, eliminating the need for quotas.

Weighting. After data collection was complete, standard procedures were followed to apply a weighting adjustment to each respondent so that the samples in each country are more representative [12]. The variables considered in weighting appear in the Appendix, Table 3. This weighting is reflected in the data shared in Section 4.

Research Objective and Data. In this paper we focus on the following research objective: How do people believe AI will affect privacy in the future? Specifically, we present emergent themes, descriptive statistics, and illustrative quotes for the following open-ended question about AI and privacy:

> *'Now we would like to ask you to think about Artificial Intelligence (AI) and privacy. In what ways will Artificial Intelligence (AI) affect privacy in the future? Please be specific.'*

This question and the four other closed-form questions we use in our analysis are provided in the Appendix, Section 7.

Coding and Analysis of Open-Ended Responses. As we reviewed responses from all countries, we iteratively refined a codebook built in previous rounds, based on emergent themes [11]. The final codebook has 368 codes on topics such as examples of AI or sentiment towards AI, 36 of which focus on privacy specifically. Any code, and multiple codes, can be assigned to a response to any open-ended question. For example, a response to the privacy question might include a code for home assistants as well as a code for hacking.

The open-ended responses were coded in the source language by Ipsos' dedicated coding team or one of their third party coding vendors. As described in McDonald et al., a variety of different approaches may be employed to improve the reliability of qualitative analysis [69]. In our case, following best practices in public opinion research for coding against multiple languages, we used professional coders, followed an iterative process to continuously improve the codes, and performed a series of hierarchical quality checks. While coders were specialized by language, they worked together to ensure consistency, sharing notes in specialized coding software. We performed multiple levels of quality checks on the resulting coding, randomly sampling from all responses in each country as well as checking all instances of select codes. In the final round, a researcher checked 10% of all responses; for the privacy question the researcher was in full agreement with all codes for 88% of the sampled responses, and the researcher was not in agreement with one or more codes for 12% of the sampled responses (range 6% to 15% across the countries) and noted that the disagreements often related to subtle coding distinctions that seemed unlikely to substantially affect broader analysis. For the privacy question, 9,765 respondents provided an answer,[4] which totaled a complete corpus of just over 100,000 words, with an average of 10.2 words per re-

[4] All respondents were required to enter text in this field, except in the United States which uses a panel that does not require responses.

sponse. Those responses were then assigned a total of 24,100 codes, an average of 2.5 codes per answer.

We used an inductive approach to explore emerging themes and common patterns in the data [33]. After the codes were assigned and we reviewed the open-ended verbatim responses in detail, four thematic groups of codes (identified separately by two different researchers) emerged as common and semantically distinct: **Data at Risk**, **Highly Personal**, **Without Consent**, and **State and Surveillance**. For example, **Data at Risk** encompassed codes such as 'Collection,' 'Available,' and 'Hacking.' We assigned each of the 368 codes to exactly one of these four thematic groups, or to a negative privacy sentiment group, a positive privacy sentiment group, an 'other privacy' group, a 'don't know' group, or an 'unrelated' group which covered a long tail of non-privacy related comments e.g., "AI causes job loss". Based on the codes that each response had been assigned, each response was assigned to one or more of these groups – for example, if a response had been assigned the code 'Hacking' and the code 'Unaware,' that response was part of the privacy groups **Data at Risk** and **Without Consent**. We summarize group/code assignment in the Appendix, Table 5.

Quantitative Analysis. While our work focuses on a thematic analysis of open-ended responses related to privacy concerns surrounding AI, we support our findings with survey statistics and modeling where appropriate. We use a χ^2 test for assessing statistical significance for survey responses involving unranked, categorical data (e.g., where a valid response may include "Don't know"). When comparing multiple distributions, we use an omnibus χ^2 test, following by pairwise χ^2 tests with a Bonferroni correction. For all models, we use a binomial distribution $Y_i \sim B(n_i, \pi_i)$ using a logarithmic link function. We report complete model odds and p-values in the Appendix for all our analysis. All calculations use the weighting adjustments of individual responses.

3.4 Limitations

We note several limitations of our methodology that should be considered when interpreting this work. First, it carries with it the standard issues attendant with survey methodology, such as the risk of respondents misunderstanding questions, poor quality translation, or respondents satisficing [47] or plagiarizing open-ended responses. We have worked to minimize these risks through piloting, use of open-ended questions in conjunction with closed-form questions, use of a translation style guide and translation review, and data quality checks. Second, online panels are not representative of the general population. While we have used a high standard of currently available online panels, we caveat our findings as not representative of the general population, particularly in China, Brazil, the Philippines, and Kenya. Third, while members of the research team have experience conducting research in all markets studied, members of the team reside in Western countries. We have worked to minimize the risk of misinterpretation by collaboration and discussion with in-country partner teams but recognize that our interpretations may lack context or nuance that would have been more readily available to local residents.

Area of life	Negative impact	Positive impact	No change	Don't know
Job availability	**51%**	23%	12%	13%
Privacy	**49%**	22%	15%	14%
Personal relationships	**46%**	21%	19%	14%
Income equality	**37%**	23%	21%	19%
Job quality	26%	**40%**	15%	18%
Creativity	25%	**47%**	13%	15%
Environmental sustainability	18%	**45%**	17%	19%
Education	15%	**52%**	17%	16%
Quality of life	15%	**52%**	16%	17%
Healthcare	10%	**59%**	16%	15%
Transportation	7%	**64%**	14%	14%

Table 1: Ranking of which areas of life people believe AI will have the largest impact on, sorted by negative sentiment. Privacy was the second highest concern for respondents, after job loss.

4 Results

We find that privacy is one of the top-most negative expectations of how AI may impact the future. In Section 4.1 we detail the strength of these concerns and explain our modeled results. Grounded in this understanding, we explore four dominant themes that underpin respondent beliefs around privacy and AI in Section 4.2. We explore solutions respondents suggested for addressing these concerns in Section 4.3.

4.1 Privacy as a Top Concern

Across the areas of life where AI may have a transformative impact in the next ten years—either positive or negative—privacy ranked as the second highest source of concern, after job loss (Table 1). In all, 49% of respondents said they expect "less privacy" due to AI.[5] While respondents recognized the potential benefits of AI—such as improving transportation, healthcare, overall quality of life, and education—our results highlight how respondents are nevertheless concerned with how AI advancements will impact their privacy.

[5]The omnibus variations between these areas of life are statistically significant ($\chi^2(30) = 17{,}822.47$, $p < .001$), as are all pairwise comparisons (all $p < .001$).

Zooming in further, we modeled how belief that AI will result in "less privacy" in the future correlates with factors such as a respondent's age, gender, and education. Here, we binarized our four answer options, treating "Less privacy" as positive samples, while treating the other three options (e.g., "More privacy", "No change", and "Don't know") as negative samples. We also controlled for various AI understanding variables including closed-form questions on how respondents define AI and how much they have heard about AI. We exclude respondents who answered "Prefer not to say" for any demographics, leaving N=9,867. We discuss our statistically significant model results below. See the Appendix, Table 6 for full modeling results.

Influence of demographics. We find that after controlling for all other factors—such as geography and AI understanding—people who are 65+ have higher odds (1.53, $p < 0.001$) of believing that AI will negatively impact privacy compared to those who are 16–24. This suggests that the experiences of, or the narratives exposed to, younger audiences may differ from older audiences. Education also has a statistically significant influence, with a higher education attainment (Bachelor's degree or more) correlating with higher odds (1.59, $p < 0.001$) of expectation that AI will negatively impact privacy compared to a lower education attainment (some primary or secondary education). We did not observe any statistically significant variations among genders.

Influence of AI understanding. Apart from demographics, a variety of dimensions for AI understanding correlate with increasing perception that AI will negatively impact privacy. As part of our quality checks, we asked respondents "Which of the following best describes Artificial Intelligence (AI)?" and provided six closed-form responses. Respondents who select "Technology that can learn or think" have much higher odds of worrying privacy will negatively impact privacy (3.13, $p < 0.001$) compared to an answer of "Not sure". Similarly, respondents who select "Self-driving car" as the "best example of Artificial Intelligence (AI)" have higher odds of privacy concerns (2.27, $p < 0.001$) compared to a selection of "Spreadsheet". Combined, both results highlight how AI understanding correlates with elevated privacy concerns, indicating that privacy concerns are not a default choice. A complete summary, by country, of the knowledge question results can be found in the Appendix, Table 4.

Exposure to news articles and narratives from peers also has a statistically significant correlation with privacy concerns. We asked respondents "In the past 12 months, how much have you heard about Artificial Intelligence (AI)?", with options ranging from "Nothing at all" to "A great amount". Respondents who select "A great amount" have lower odds of privacy concerns (0.76, $p < 0.001$) compared to those who hear "a moderate amount". Conversely, respondents who select "A little bit" have higher odds (1.23, $p = 0.001$). This suggests that cursory exposure to AI narratives correlates with elevated privacy concerns, whereas people with broad exposure to AI narratives (potentially due to personal interest) may be more excited by the possibilities that AI might achieve.

Influence of geography. We find that after controlling for demographics and understandings of AI, the United States has the strongest belief that AI will negatively effect privacy, which we treat as a baseline for modeling. In terms of odds, our remaining countries rank as follows: Germany (0.61), Australia (0.61), Brazil (0.52), South Korea (0.50), the Philippines (0.48), Kenya (0.47), China (0.42), Russia (0.40), and Japan (0.24), all with $p < 0.001$. As such, the United States represents an outlier where respondents have strong privacy concerns surrounding AI, while China, Russia, and Japan are outliers where respondents have lower privacy concerns.

4.2 Privacy Themes

We investigated respondents' expectations regarding AI and privacy by analyzing their open-ended responses. Table 2 shows the prevalence of each theme by country, with **Highly Personal** being the most common at a 31% global average and **Without Consent** being the least common of our themes at a 5% global average. In total, 58% of respondents touched on one of our four themes, or generally expressed a negative expectation. Even if they did not express one of our themes, many respondents expect that AI will have a negative effect on privacy, or even inevitably lead its complete dissolution. Some said the deterioration of privacy due to AI is already (far) underway and will only get worse over time. While negative sentiments were predominant, 12% of respondents expressed that AI could be positive for privacy. While we do not include analysis of positive expectations in the paper, we include a sampling for the interested reader in the Appendix, Section 8.

> There will be less privacy, of course. *–Russia*[6]
>
> It is likely to adversely affect privacy *–South Korea*
>
> I believe that every day our privacy will be increasingly invaded, until the time comes when we will have no more privacy *–Brazil*
>
> It will wreck privacy *–Australia*

Overall, most respondents—75%—shared relevant comments on the state of privacy and AI. The remaining 25% of responses included 'don't know' or responses which only had codes that seemed unrelated to privacy. For the remainder of this section, we focus on our four major themes that explain why respondents feel AI will lead to less privacy in the future.

[6]Throughout the paper, we share complete verbatim responses (in some cases translated). In some cases we have made minor edits for readability, e.g., to correct typos or grammatical errors.

Privacy themes, from open-ended coding	Global average	Australia	Brazil	China	Germany	Japan	Kenya	Philippines	Russia	South Korea	United States
Highly Personal	31%	33%	32%	33%	17%	26%	43%	45%	20%	26%	42%
Data at Risk	29%	35%	29%	29%	15%	26%	44%	41%	13%	22%	35%
State and Surveillance	12%	20%	13%	8%	15%	6%	14%	9%	8%	10%	25%
Without Consent	5%	6%	4%	5%	2%	6%	7%	7%	1%	3%	8%
Privacy sentiment, from open-ended coding											
Negative sentiment or part of a theme	58%	66%	57%	58%	45%	46%	75%	68%	42%	54%	77%
Positive sentiment	12%	8%	11%	17%	8%	4%	23%	11%	11%	17%	4%
Overall response quality											
Expressed any privacy statement	75%	76%	73%	81%	67%	64%	94%	81%	61%	69%	82%
Don't know	18%	20%	17%	11%	25%	30%	3%	10%	21%	24%	16%
Any other response	8%	4%	10%	8%	8%	6%	4%	9%	18%	7%	2%

Table 2: Breakdown of privacy themes across countries. Themes do not add up to 100% due to the possibility of zero or multiple themes per response. We also report the aggregate frequency of responses that fit into a theme along with other negative leaning responses which did not fit into a theme and for comparison responses which showed positive sentiment towards privacy. The final section shows overall response quality, which does sum to 100%, consisting of privacy-related responses, 'don't know' responses, and a small number of responses that were assigned codes that seemed unrelated to privacy.

4.2.1 Data at Risk

Respondents believe AI increases privacy risks due to the *scale of data collection*. They expressed concern that because AI requires data to work, more data will be gathered; it will be collected from more devices, many of which are networked; it will then be aggregated and linked together, potentially across products and surfaces; and data and inferences will then be available in online databases and servers. Our respondents felt this accumulation increased the risks to their data, as it could more easily be accessed or misused, or it could leak or be breached by hackers. This theme is most prevalent in the Philippines and Kenya and less prevalent in Russia and Germany (Table 2). Concern is also higher among younger people ages 16–24 (odds = 1.91, $p < 0.001$). See the Appendix, Table 7 for full modeling results.

AI needs data. Respondents observed that AI needs data in order to learn and operate, and that the more information it gets about people, the more efficient and accurate it will become. The view that AI inevitably encourages the accumulation of large amounts of data led to the conclusion that AI is negative for privacy.

> I don't think there will be much privacy, because artificial intelligence needs a lot of data and information to work! –*Brazil*
>
> AI requires a lot of human data –*China*
>
> For machines to think, they need to analyse and base their decisions on data. Data will be a hot commodity and companies will look for all ways for you to give them your data to use as inputs for AI –*Kenya*
>
> The collection of personal data and information is one of the fundamentals of artificial intelligence. For this reason, I believe there will be a negative impact on users' privacy. –*Brazil*
>
> Our data will be constantly collected, even more so than it is now, to feed machine systems. Nothing will be private or sacred anymore. –*Australia*

Multiple, connected sources. Respondents spoke of increasingly expansive data collection and user tracking across smartphones, computers, smart home devices, IoT devices, self-driving cars, and more. They described data being constantly extracted from devices, often highlighting that network connectivity facilitates data collection and increases personal exposure. Some also observed that AI gathers data from the internet or social media.

> It's already here. Every time I use my phone or the internet, AI is at work gathering all information passing through my devices –*Philippines*
>
> It makes me think I shouldn't use any connected devices as AI will know exactly what I'm doing at all times. I don't like that at all –*United States*
>
> With everything interconnected privacy will not exist –*Brazil*
>
> there will be no privacy - the AI will have access to ALL information –*Russia*

Crosslinked and aggregated. Beyond tracking across different devices, AI-related data collection and cross-linking was

seen as occurring across different services and systems, and some respondents suggested that different companies or organizations share data with each other, perhaps indiscriminately. This contributed to a sense that a growing amount of personal information is flowing together.

> AI could easily connect seemingly innocuous information from across the internet to gain insight into the lives of people and reveal things they might want to keep to themselves *–United States*

> Facial recognition and other ways to track people will negatively affect privacy. AI will allow a lot more information to be gathered and collated. *–Australia*

> Everything about us will be collected and placed on one platform that can easily be accessed by governments or advertising agencies *–Kenya*

Available and vulnerable. Respondents felt troves of personal data, once accumulated, were available online and vulnerable to legitimate or illegitimate misuse, mishaps, and more. Cloud storage, internet connectivity, or data being held in multiple places were seen as increasing exposure. Respondents described multiple actors who posed a risk to this data. Some might have legitimate but still problematic access, e.g., companies, governments, or wealthy and powerful people who control AI were often characterized as suspicious or bad actors, and respondents observed that the concentration and availability of large amounts of information might provoke unethical use, particularly given lack of strong regulations. Respondents also pointed out that both system errors and human mismanagement of data can compromise security, and even inadvertent leaks can make data completely public.

> I'm afraid that there will be some glitch in the system and all my information will be open to a stranger *–Russia*

> People will trust AI with more and more personal information and there could be a big leak of that data into the open space of the Internet *–Russia*

Cyberattacks, hackers. Beyond those with legitimate access, others might gain access through illegitimate means. AI was seen as prone to cyberattacks, hackers, criminals, malfunction, and more. Potential data breaches or leaks were particularly concerning because AI was seen as having such a large amount of personal data. Respondents were worried that hackers would be able to take advantage of vulnerabilities and security holes to steal their confidential information, or to take over devices to spy on them. Respondents said that even if careful protective measures were taken, safety against hackers was not guaranteed.

> If it's in a computer it can be hacked *–Philippines*

> all personal data will be input to a cloud system that can possibly be hacked by an advanced person *–Philippines*

> More and more our data will be in databases exposed to strong intrusions by increasingly skilled hackers. *–Brazil*

> There will be no privacy since access of personal information will be easy. I also do not think that artificial intelligence can prevent hackers from accessing information. *–Kenya*

> No matter how secure it is, I think it will make it easier for leakage of personal information to occur. *–Japan*

> I think AI will have access to all our personal data and I believe that there is always a way for malicious people to circumvent security. *–Brazil*

4.2.2 Highly Personal

Beyond scale, respondents characterized how AI increases risks due to *sensitivity of collected data and derived insights*. They pointed out that as AI advances, it becomes better at finding out about people's private lives, AI's data and inferences can be highly personal, and these personal insights are leveraged to influence decisions and behavior. This theme is prominent for people from the United States, Kenya, and the Philippines; and less prevalent in Germany and Russia (Table 2). Concern is higher among younger people ages 16–24 (odds $= 1.51, p < 0.001$). See the Appendix, Table 8 for full modeling results.

Personal data. Respondents described a wide range of sensitive or confidential data that is gathered about people: financial, health, relationships, social media and internet history, entertainment, hobbies, education, occupation, demographic data such as race and religion, household activities, location, and more. There was a strong sense of intrusion and loss of privacy, with the sentiment that this highly detailed data penetrates all aspects of life, seems like more information than necessary, and may be more than people want to share. The view was expressed that people's entire lives will be documented, and they will become entirely "transparent" since everything about them will be known.

> Much more private data will be collected and used. Contacts, places you have been to, people you call, websites you visit, what books and articles you read, products you buy and use. Tracking via face recognition. Masses of information to be used to predict behaviour. *–Australia*

> Collecting our words and actions down to the smallest detail – *South Korea*

> will brazenly violate personal life *–Russia*

> I think it will affect privacy in that AI will be privy to immense amounts of our personal data which we do not even realise is available...e.g. doctors' records on us, school records, tax account records, etc. *–Australia*

Personal insights. Respondents explained that this data is combined and processed to yield personal insights. For example, AI was seen as leveraging large data sets to learn people's tastes and interests, surface behaviors and habits, predict future actions, create profiles, or infer feelings or personality traits. Respondents highlighted that such use of big data leads directly to loss of privacy. They also shared that AI may infer

things that people would prefer to keep private, or perhaps reach profound insights about them that they are not even aware of themselves. A few mentioned that AI may have a superhuman capacity to draw conclusions or even look into people's thoughts and read their minds.[7]

> Big data will reveal you *–China*
>
> Our privacy will be totally affected because technologies will capture data about our conversations, consumption habits, medication, contacts and everything else to create a database and predict things that may interest us *–Brazil*
>
> Privacy is violated, any citizen can be tracked, their psychological portrait can be created, conclusions can be drawn about their character, etc. *–Russia*
>
> Accumulation and analysis of privacy information. Before you know it, you will learn things that you don't know about yourself. *–Japan*
>
> AI will be able to predict all human individuals' decisions in society. It is connected to devices that are always listening and watching us and will have access to every electronic communication or record that we have ever had. It will know us better than any human possibly could. *–United States*

Influencing decisions and behavior. Companies, governments, and powerful people were seen as using these personal insights for profit or other motives. Information was seen as a tool to influence or manipulate people's decisions and behavior, purchasing and otherwise. Some saw AI as controlled by and benefiting the wealthy, and expected that those with limited means would have more difficulty maintaining their privacy than those with substantial means. Respondents also spoke of companies pushing people towards consumption with targeted advertising and customized services, and sentiment towards companies' use of AI was often extremely negative. We discuss government use of information in more detail below.

> It can be used in an evil way by powerful groups in order to take advantage of the personal data of the population, to manipulate and control them *–Brazil*
>
> In the future, people will have no privacy, and any personal data will be controlled by a few people or the government *–China*
>
> All data will be stored in the "Cloud" on which people with power and technology will be able to access it at any given time. *–Philippines*
>
> Large corporations will ruthlessly use AI to market their products or services to a wider demographic by sharing clients' private information with each other. *–Australia*
>
> it will be impossible to resist the advertising, it will be very personalized and literally force you to make the decisions the advertiser wants *–Russia*

[7] If a sentiment was rare and did not occur robustly in the data, we note that it was expressed by "a few" respondents.

4.2.3 Without Consent

Apart from scale and sensitivity, respondents expressed concern that people do not give meaningful *consent* because they are not asked and may not even be aware of how their data is used or gathered, and also because users of online services are required to provide personal data in order to use AI services. Of our themes, this is the least prevalent across countries as shown in Table 2. See the Appendix, Table 9 for full modeling results.

Data gathering and use occur without consent. Respondents expressed concern that AI-powered products and systems gather and use data without people's consent or authorization. Some called out AI's ability to make predictions or draw inferences about non-disclosed aspects of people's lives, without their permission. Others emphasized that once AI gains access to data, people have no control over how it is used or shared.

> People's data will be invaded whether they know and give their consent or not. *–Kenya*
>
> Invasion of privacy by spying on me without my consent – *South Korea*
>
> Am not sure I will feel safe in a society where even nanny cams, smart tvs and others will collect personal data without my consent. *–Kenya*
>
> The vast amounts of data now possessed or readily available will be even more searched and analyzed to predict any one individual's patterns and tendencies. The individual has very little meaningful control over how that will be used to influence them or society. *–United States*

Unaware. AI was also seen as operating without people's knowledge. As seen in Section 4.2.4, activities such as spying were particularly likely to be called out as occurring without people being aware. But beyond that, respondents expressed concern that people would not know what AI systems knew about them, what inferences had been drawn, when data was being gathered, whether their information had been stolen, or when or how AI was being used. This lack of information was seen as concerning not only because it compromised trust and transparency, but also because it compromised people's ability to directly manage and control their privacy. Some respondents suggested that some people are more savvy about technology than others, and therefore better able to protect themselves from possible AI-related privacy infringements, and emphasized that those who are less aware of privacy can not protect themselves effectively and will be disproportionately negatively impacted.

> It has already invaded households beyond what the majority of people know. There is no privacy now. *–United States*
>
> We will not have privacy anymore. Companies will use our data and we won't even know. *–Brazil*
>
> The public is deceived and privacy continues to be violated behind the scenes. *–Japan*

Personal information required to use services. As has been observed in other contexts [15, 16], respondents felt they were required to provide personal information in order to access services and participate in modern society. For example, some said personal or identifying details are required to register for AI-powered websites, services, and products. This was viewed transactionally, that users provide private information to access services, and further, provide larger amounts of information to get the personalized services and efficiency that AI offers. This contributed to a sense that individuals must give up privacy to improve algorithmic decisions and gain convenience.

> We will be forced to give up more private information or will not be able to use new products or systems *–Australia*

> Useful features will be available in exchange for the disclosure of personal information. *–Japan*

> Artificial intelligence requires people to reveal themselves, while inevitably exposing their privacy *–China*

> We lose some privacy in exchange for more efficiency. *–Philippines*

4.2.4 Surveillance and State

Independent of how AI obtains data, respondents shared concerns about how AI can conduct constant surveillance and can be used by governments to fight crime or for population control. This theme is most prevalent for people from Australia, Germany, and the United States as shown in Table 2. This theme is also more popular among men (odds $= 1.43, p < 0.001$). Conversely, this theme is less prevalent in Japan (odds $= 0.24, p < 0.001$), China (odds $= 0.31, p < 0.001$), and Russia (odds $= 0.35, p < 0.001$). See the Appendix, Table 10 for full modeling results.

AI conducts surveillance. AI was often described as an instrument of surveillance. The sense of being surveilled made some respondents feel strange, creepy, or that they had nowhere to hide, or even that they were naked or in a "glass house". Voice assistants and smart home devices such as Siri and Alexa were highlighted as listening devices that collect information, and AI was characterized as surreptitious, for example, spying, eavesdropping, or watching covertly. Sometimes it was explicitly called out as taking these actions without consent. AI was further described as constantly operating, recording, and analyzing people's every move, which contributed to respondents' sense of being continuously monitored and evaluated.

> AI will be the ultimate spy *–Australia*

> I feel like I'm being monitored at all times. *–Japan*

> An AI is like a device with eyes and ears that is watching you 24/7, and storing your personal information *–Philippines*

> you won't know who or what is watching *–Germany*

> We have become a surveillance society and privacy is no more. *–Japan*

AI is omnipresent. Respondents also called out AI's ubiquitous nature, often describing specific devices or locations which contribute to the sense that AI can be all-seeing and all-knowing. For example, respondents mentioned increased prevalence of cameras (CCTV and otherwise), proliferation of electronic devices, drones overhead, and the watchful eyes of robots that observe and evaluate people. They spoke of being monitored at home, in public spaces, on public transportation, in the car (e.g. self-driving Ubers with cameras), and more generally, "everywhere you go", as well as during all online activities.

> It will be everywhere and in everything we use, being able to monitor us *–Brazil*

> You cannot dress freely at home, in case AI is out of control *–China*

> If everything is artificial, people will be afraid to go to the bathroom and out of the blue the toilet will turn a robot or whatever. *–Brazil*

> Surveillance cameras are located everywhere, AI will be able to find any person everywhere and monitor their entire path and actions. *–Russia*

AI identifies people. Beyond this type of monitoring, AI's ability to identify people through mechanisms such as facial recognition was called out as a key enabler of increased surveillance, and correspondingly AI was seen as reducing people's ability to be anonymous. Accordingly, AI was viewed as making it easier for governments to manage and evaluate citizens. While effects such as improved policing and criminal investigations were seen as beneficial, facilitating greater access to personal information by law enforcement and security agencies was raised as a concern.

> Facial recognition makes it much easier to follow individuals and spy on them. That is good when looking at crime but it is very different when it comes to people going about their normal legal life *–Australia*

> I think that even faster and more accurate identification of individuals is progressing, and in some cases, I think that constant observation is also possible. *–Japan*

> The government can find out the identity of any individual without having their permission. *–Philippines*

> AI makes it easier to find a human being in all the data chaos *–Germany*

AI serves state purposes. Beyond use for law enforcement, AI was seen as serving state purposes such as government control, and was associated with a police state or surveillance state. In fact, some suggested that law enforcement was a pretext to gather data for other government purposes. Regardless, use of AI for state purposes was generally viewed negatively, and respondents across a wide range of countries positioned AI as a potential tool of government oppression, propaganda, or human rights violations.

> The use of machines to determine a person's risk to the country will be a breach of one's privacy. *–Kenya*

> Data about all citizens will be collected, everybody will be under the state's microscope *–Russia*

> It will be possible to spy on the population even more easily than it is now. It will be even easier to control our lives. *–Germany*

> In dictatorships, artificial intelligence can be used to identify potential political crimes, resulting in violations of freedom of conscience. *–Japan*

> There will be no privacy. They're going to use AI to predict and control the behavior of the population. *–Brazil*

Respondents connected AI with existing real-world and fictional examples of government surveillance and control, such as the Chinese Social Credit System and Big Brother in George Orwell's 1984. Respondents expressed concern that AI would bring these scenarios to fruition in their own countries.

> Artificial intelligence will make it easier for governments and companies to monitor the population in a more aggressive way. The personal credit system deployed in China and which has been gaining ground in other countries is an example of this. *–Brazil*

> I feel there will be little to no personal privacy, and that worries me. I believe that the Orwellian world will become more reality than fiction. *–Australia*

> As in Orwell's book, the more technology, the more observation, and the more exposure, the less privacy *–Brazil*

4.3 Solutions

Respondents said it is important to take steps to alleviate privacy concerns with AI, for example by pursuing responsible development, regulation, the development of new privacy and security technologies, or setting expectations that end users will manage their own privacy. While these ideas were expressed less frequently than our four main privacy themes, we share them here to provide insight into public attitudes regarding potential improvements.

Responsible development. Respondents observed that the impact of AI on privacy depends on the choices and moral character of the people who design, build, and deploy it. Some alluded to principles of responsible development, expressing optimism that careful design and strict security measures can mitigate privacy risk. On the other hand, others called out companies and governments as untrustworthy or unethical, e.g., raising concerns that companies would make questionable choices to maximize profit, that organizations might not be competent to execute well-intended protection plans, or that governments do not have a favorable historic track record for handling sensitive information. Open questions regarding responsible development left AI's expected future impact on privacy uncertain, but respondents felt one way or another AI would have a big impact on privacy.

> I don't think it will affect privacy in a negative way if it is designed correctly *–Australia*

> I think it's not so much AI itself, but how data stored by AI is handled. *–Japan*

> It's not that I don't trust AI, it's that I can't trust the humans in charge of it. *–South Korea*

Regulation. While some respondents focused on responsible development (which is sometimes associated with self-regulation, although it can also occur within more formal legal frameworks), others focused more directly on formal regulatory measures. Some believed that protective laws are already in place in their countries, while others expressed concern that currently there are no guardrails and said such laws urgently need to be developed. Some were optimistic that regulatory protection would be sufficient while others expressed concern that its effectiveness would depend on the values and priorities of the government, or concern that regulatory response will lag development and deployment of new AI technologies.

> Nowadays artificial intelligence is already invasive. I believe that in the future it will worsen if the authorities do not have greater control. *–Brazil*

> Without the right protections AI will be able to obtain sensitive data in ways that currently don't exist. This will require new laws to be created to protect privacy in ways that have not been considered to date. *–Australia*

> My opinion is that AI will cause loss of privacy if the rules and regulations are not properly managed. *–Kenya*

> I think that technology will develop in the future, but privacy protection measures or laws will be stronger. In other words, the state will control AI in terms of privacy. So I think what happens to us now, will happen in the future in terms of privacy. *–China*

> Bad actors will use it for morally dubious purposes. Some will use it to improve lives. Our laws will take decades to catch up to the technology to appropriately regulate it. *–United States*

Advanced protective technology. Respondents sometimes framed AI technology and privacy/security technology as opposing forces, and spoke of the need to develop new privacy/security measures to keep pace with new threats posed by AI. While it has long been a desire of the Privacy Enhancing Technology community to develop useful, usable technologies that help people protect their privacy, progress has been limited [44, 83].

> There is a war between a robot that steals and a robot that tries to protect. *–South Korea*

> The technology to avoid exposure to privacy and the technology to acquire private information are developing at the same time *–South Korea*

> Perhaps a lot more user data will be collected hence a higher risk to exposure in the case of hacking incidents. Hence, cyber security will need to be top notch. *–Kenya*

Individual action. While the predominant attitude was that companies and governments should work to protect users' information, for example, via responsible development or regulation, a few respondents did value individual action especially in combination with end user privacy controls. A few mentioned that while individuals need to manage their own privacy in theory, some people do not have the information or tech savvy to do so, which will lead to privacy exposure.

5 Discussion

Here we show how the themes we identified come together to build one common, overarching narrative of how our respondents believe AI will shape the future of privacy. We discuss ways the research community, regulators, and technologists can consider mitigating these privacy issues for AI systems, and conclude with further suggestions for engaging the public.

5.1 Overarching Narrative

In working with the data, a dominant, interconnected narrative emerged. While most respondents did not cover all aspects of this narrative, many of them spoke to one or more pieces of it. This narrative encompasses our main themes as well as specific ideas they are composed of.[8] In this overarching narrative, many ideas are causally connected, e.g. AI needs data, therefore AI involves creating a large dataset, which is then at risk from hackers. To illustrate this narrative, we created the following composite consistent with the content, language, and tone of responses we received across countries [35, 97].

> *Data is the foundation of AI, so it involves gathering massive troves of data from cameras, smartphones, home assistants, self-driving cars, robots, social media, and many other connected devices and products that touch all aspects of people's lives. Often this data is collected surreptitiously or without consent, and AI can conduct constant surveillance, identifying people and tracking their movements and activities with technologies like facial recognition. AI combines and analyzes all this data to draw highly personal or even invasive conclusions about individuals, which can be used to influence or manipulate their decisions and behavior, purchasing or otherwise. Between data and inferences, AI may learn essentially everything about a person. All this personal information sits around online or in the cloud where it is at risk from hackers and malfunction. Companies, governments, and powerful people control AI and can use it for good or bad purposes. Companies typically use it for profit and governments typically use it to fight crime or control the population. Because of AI's nature and capabilities, its use leads to substantial or even total loss of privacy.* –Composite Across Respondents and Countries

This narrative appears across all countries in our sample, with varying emphasis and some local twists (e.g., elevated antagonism towards corporate marketing in the United States, or particular emphasis on government surveillance in Russia). Similarly, many individual respondents touched on various combinations of these ideas, often calling out two or three (or more) ideas from this overarching narrative (one common pattern was to connect personal data and hackers). For example, here is a particularly long response:

> AI will become more intrusive and we will continue to lose the last bits of privacy we have if we don't enact laws to restrict how it is used. The most obvious example is more cameras will be installed in all public places, using AI to process the images/video for various reasons such as safety (criminal "behaviour"), access to places, identification verification, etc. London already has a network like this so they have already lost any privacy in public. Technology already tracks where we go via our phones that are ubiquitous, AI will continue to expand to use that information along with previous behaviour that is stored to do things like show us "personalized" advertising. The data will be sold to other companies to use with their proprietary AI that will be used to evaluate people for jobs, loans, housing, etc. Basically, data collection will increase and AI will be developed to connect a lot of disparate information to personally identify us and then AI will be used to influence important decisions about our lives - and we won't even know it. A combination of super data collection and AI will be the death of privacy in the future. *–United States*

Beliefs about technology are often grounded in folk models, with inconsistent or inaccurate elements. By contrast, this narrative and its language are well-aligned with messages in the popular press, e.g., [3, 17, 30, 46, 48, 55, 64, 67, 70, 72, 84, 99] as well as expert opinion expressed in policy briefs [34, 58] and scholarly articles [2, 29, 31, 60, 66, 76, 101]. Further, it is largely consistent with common factual representations, and does not appear to contradict itself. While expressions of ideas were sometimes hazy or incomplete, respondents across countries largely seem to be discussing pieces of the same coherent narrative, rather than expressing completely different ideas. An area for future work is to investigate how specifically respondents came to have these beliefs. Further, these beliefs merit further study as AI becomes more common and as the most visible examples or messages in the press shift. As an example, the recent rise of generative AI technologies (e.g., ChatGPT, Midjourney) may or may not lead respondents in future surveys to have different privacy considerations.

5.2 Mitigating Concerns

Addressing the four privacy themes that we observed in our study will require a unique combination of education, technol-

[8] From an analytic perspective, specific ideas in the narrative generally correspond to codes in our analysis, e.g. AI needs data, AI listens, data is available, and each of these codes is assigned to one of our themes. Solutions are a logical extension of the main narrative, and while less common, respondents sometimes included them along with other ideas from the narrative. Ideas that were positive about AI's expected impact on privacy generally seemed separate and did not tend to co-occur with the overarching narrative.

ogy design, and policy changes.[9] We describe an initial set of potential directions for researchers, platforms, and policy makers to help mitigate user concerns.

Highly Personal. One potential direction to address the sensitivity of data *ingested* by AI models would be to leverage privacy enhancing learning algorithms. These strategies—such as student-teacher models [79], federated learning [62], and differential privacy [53]—help to ensure that models do not memorize an individual's sensitive training data, which might otherwise be leaked depending on the model's architecture [21, 22]. However, while these strategies may add protection in some cases, they may not be suitable or effective in other cases. For example, these strategies do not address user concerns that AI systems can be used to *infer* sensitive attributes; or indeed, surface inferences about an individual they might otherwise have thought private or idiosyncratic [45]. While the Overton window around acceptable AI applications is likely to shift in the next few years (e.g., due to benefits of new AI technologies), commitments around AI principles from platforms and potential privacy regulation can help to assuage concerns that technical solutions are presently unable to address.

Data at Risk and Without Consent. Addressing privacy concerns around data collection for AI algorithms and consent is more challenging. These concerns dovetail long-standing user sentiment that platforms monitor every transaction, interaction, or click for advertising and recommendation algorithms [42, 91, 94, 105] and are thus likely only to be exacerbated by emerging AI technologies. Policies such as the EU General Data Protection Regulation (GDPR)[10] have attempted to ensure any data collection that occurs has a pre-defined use case, and requires "freely given, specific, informed and unambiguous" consent, which users must be able to withdraw. Policy makers might explore similar applications to AI training data. Concerns around inadvertent exposure are easier to address: techniques like federated learning [62] represent a promising direction to ensure that non-aggregated data never leaves a user's device, thus providing some mitigation against data breaches or insider risk [80].

State and Surveillance. Addressing potentially harmful applications of AI—particularly those operated by government or quasi-government actors—remains an open challenge. Platforms can help to prevent state surveillance by committing to responsible practices that constrain the use or distribution of certain technologies, or even prohibit their development entirely [54][11]. Researchers have considered adversarial techniques to deceive facial recognition [86] and audio tracking [23], as well as jamming data collection entirely [31], but these remain proof-of-concept only. Further research is needed into policy and technical mitigation of AI-assisted surveillance.

[9] While we do see some differences based on age, education, understanding of AI, and country on overall attitudes regarding the impact of AI on privacy, we see the same four themes arising across the countries studied. Therefore, we believe that when we consider mitigations, we can take a global perspective.

[10] https://eugdpr.org/

[11] https://ai.google/principles, https://www.ibm.com/artificial-intelligence/ethics, https://www.microsoft.com/en-us/ai/our-approach

5.3 Civic Participation

Experts and members of the public have called for greater public participation in policymaking and decisions about AI [101]. Such civic engagement can encompass a wide range of activities, from attending city council meetings to express opinions about whether local law enforcement should use facial recognition, to voting for laws or candidates aligned with one's own beliefs about the use and development of AI, or participating in joint problem-solving with policy makers and technologists.

However, public knowledge has been viewed as a significant barrier to such participation for many aspects of AI such as explainability and automated decision-making [77, 88, 100, 101], sometimes addressed through small-scale interventions in which members of the public receive training in order to provide feedback on a policy question [7, 9, 50, 93]. Happily, our research is cause for optimism that the public may be better prepared than expected to discuss privacy-related aspects of AI. While some members of the public may not have a full general understanding of AI and many may not have a detailed understanding of its specific operations, many members of the public do appear to be conversant in high level-issues and have well-described concerns regarding AI's impact on privacy, and these attitudes are broadly aligned with issues raised by experts. Our findings are encouraging for both immediate public participation and facilitated joint problem-solving.

6 Conclusion

In this work, we surveyed 10,011 respondents in 10 countries to understand how and why people believe AI will affect privacy in the future. We found that privacy was a consistent, global concern, with 49% of respondents saying they would have "less privacy" due to AI over the next 10 years. We presented a thematic analysis of privacy concerns surrounding AI and identified four key themes, which align with experts and privacy advocates. These themes struck on how the substantial data required to train AI models may be misused or hacked; how data and inferences may reveal highly personal details; that data collection and use can occur without meaningful consent; and that AI may be used for surveillance or government purposes. We discussed avenues that researchers, industry, and policy makers might explore to mitigate these concerns, such as adopting privacy enhancing technologies or AI principles. In light of the public's comprehension of the benefits and potential harms surrounding AI, discussions on the future of AI and privacy can potentially leverage civic participation to arrive at the best balance of solutions.

Acknowledgments

We thank Dan Altman, Elie Bursztein, Jen Gennai, Tushar Gupta, Angela McKay, and Ashley Walker of Google for their valuable contributions to this work. We thank Christopher Moessner, Laurie Pettigrew, and Wendy Whitfield for their expertise fielding and coding the survey. We thank the cApStAn team for their important contributions to linguistic quality.

References

[1] N. Abdi, K. M. Ramokapane, and J. M. Such. More than smart speakers: Security and privacy perceptions of smart home personal assistants. In *Fifteenth Symposium on Usable Privacy and Security (SOUPS 2019)*, pages 451–466, Santa Clara, CA, Aug. 2019. USENIX Association.

[2] N. Abdi, X. Zhan, K. M. Ramokapane, and J. Such. Privacy norms for smart home personal assistants. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2021.

[3] M. Anderson. Facebook privacy scandal explained. *CTV News*, Apr. 2019.

[4] M. Andrejevic, R. Fordyce, L. Li, and V. Trott. Australian attitudes to facial recognition: A national survey, 2019.

[5] ARM | Northstar. AI today, AI tomorrow. Awareness and anticipation of AI: A global perspective, 2017.

[6] ARM | Northstar. AI today, AI tomorrow: The Arm 2020 global AI survey, 2020.

[7] A. Armour. The citizens' jury model of public participation: a critical evaluation. In *Fairness and Competence in Citizen Participation*, pages 175–187. Springer, 1995.

[8] B. Auxier, L. Rainie, M. Anderson, A. Perrin, M. Kumar, and E. Turner. Americans and privacy: Concerned, confused and feeling lack of control over their personal information. *Pew Research Center*, November 2019.

[9] B. Balaram, T. Greenham, and J. Leonard. Artificial intelligence: Real public engagement, 2018.

[10] J. Beck. People are changing the way they use social media. *The Atlantic*, June 2018.

[11] H. Beyer and K. Holtzblatt. *Contextual Design: Defining Customer-centered Systems*. Elsevier, 1997.

[12] P. P. Biemer and S. L. Christ. Weighting survey data. In E. D. de Leeuw, J. J. Hox, and D. A. Dillman, editors, *International Handbook of Survey Methodology*, pages 317–341. Lawrence Erlbaum Associates, New York, NY, 2008.

[13] C. Bloom, J. Tan, J. Ramjohn, and L. Bauer. Self-driving cars and data collection: Privacy perceptions of networked autonomous vehicles. In *Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017)*, pages 357–375, Santa Clara, CA, July 2017. USENIX Association.

[14] Blumberg Capital. Artificial intelligence in 2019: Getting past the adoption tipping point, 2019.

[15] K. Bongard-Blanchy, A. Rossi, S. Rivas, S. Doublet, V. Koenig, and G. Lenzini. "I am definitely manipulated, even when I am aware of it. It's ridiculous!" – Dark patterns from the end-user perspective. In *Designing Interactive Systems Conference 2021*, pages 763–776, 2021.

[16] C. Bösch, B. Erb, F. Kargl, H. Kopp, and S. Pfattheicher. Tales from the dark side: Privacy dark strategies and privacy dark patterns. *Proceedings on Privacy Enhancing Technologies*, 2016(4):237–254, 2016.

[17] N. Bowles. Thermostats, locks and lights: Digital tools of domestic abuse. *The New York Times*, June 2018.

[18] E. Brynjolfsson and A. McAfee. *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*. WW Norton & Company, 2014.

[19] T. Bucher. The algorithmic imaginary: exploring the ordinary affects of Facebook algorithms. *Information, Communication & Society*, 20(1):30–44, 2017.

[20] J. Caltrider. 10 fascinating things we learned when we asked the world "How connected are you?", 2017.

[21] N. Carlini, J. Hayes, M. Nasr, M. Jagielski, V. Sehwag, F. Tramèr, B. Balle, D. Ippolito, and E. Wallace. Extracting training data from diffusion models. *arXiv preprint arXiv:2301.13188*, 2023.

[22] N. Carlini, C. Liu, Ú. Erlingsson, J. Kos, and D. Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *USENIX Security Symposium*, volume 267, 2019.

[23] N. Carlini and D. Wagner. Audio adversarial examples: Targeted attacks on speech-to-text. In *2018 IEEE Security and Privacy Workshops (SPW)*, pages 1–7. IEEE, 2018.

[24] D. Castro. The U.S. may lose the AI race because of an unchecked techno-panic. *Center for Data Innovation*, March 2019.

[25] S. Cave, K. Coughlan, and K. Dihal. "Scary Robots": Examining public responses to AI. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES 2019)*, pages 331–337, 2019.

[26] S. Cave, C. Craig, K. S. Dihal, S. Dillon, J. Montgomery, B. Singler, and L. Taylor. *Portrayals and perceptions of AI and why they matter*. The Royal Society, 2018.

[27] S. Cave, K. Dihal, and S. Dillon. *AI Narratives: A History of Imaginative Thinking about Intelligent Machines*. Oxford Scholarship Online, 2020.

[28] CBS News. 60 Minutes/Vanity Fair poll: Artificial intelligence, March 2016.

[29] G. Chalhoub, M. J. Kraemer, N. Nthala, and I. Flechais. "It did not give me an option to decline": A longitudinal analysis of the user experience of security and privacy in smart home products. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2021.

[30] B. X. Chen. Here is how to fend off a hijacking of home devices. *The New York Times*, Feb. 2017.

[31] Y. Chen, H. Li, S.-Y. Teng, S. Nagels, Z. Li, P. Lopes, B. Y. Zhao, and H. Zheng. Wearable microphone jamming. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, page 1–12, 2020.

[32] C.-H. Chuan, W.-H. Tsai, and S. Cho. Framing artificial intelligence in American newspapers. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES 2019)*, pages 339–344, 2019.

[33] V. Clarke, V. Braun, and N. Hayfield. Thematic analysis. In J. A. Smith, editor, *Qualitative psychology: A practical guide to research methods*, pages 222–248. Sage, London, third edition, 2015.

[34] L. Cranor, T. Rabin, V. Shmatikov, S. Vadhan, and D. Weitzner. Towards a privacy research roadmap for the computing community. *arXiv preprint arXiv:1604.03160*, 2016.

[35] J. W. Creswell and C. N. Poth. *Qualitative Inquiry and Research Design: Choosing among Five Approaches*. Sage Publications, fourth edition, 2018.

[36] L. Dencik and J. Cable. The advent of surveillance realism: Public opinion and activist responses to the Snowden leaks. *International Journal of Communication*, 11:763–781, 2017.

[37] M. A. DeVito, J. Birnholtz, and J. T. Hancock. Platforms, people, and perception: Using affordances to understand self-presentation on social media. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 740–754, 2017.

[38] Edelman. 2019 Edelman AI survey, March 2019.

[39] M. Eslami, A. Rickman, K. Vaccaro, A. Aleyasen, A. Vuong, K. Karahalios, K. Hamilton, and C. Sandvig. "I always assumed that I wasn't really that close to [her]": Reasoning about invisible algorithms in news feeds. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 153–162, 2015.

[40] E. Fast and E. Horvitz. Long-term trends in the public perception of artificial intelligence. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[41] C. Fiesler and B. Hallinan. "We are the product": Public reactions to online data sharing and privacy controversies in the media. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 1–13, 2018.

[42] A. Friedman, B. P. Knijnenburg, K. Vanhecke, L. Martens, and S. Berkovsky. Privacy aspects of recommender systems. *Recommender Systems Handbook*, pages 649–688, 2015.

[43] C. Funk, A. Tyson, B. Kennedy, and C. Johnson. Science and scientists held in high esteem across global publics. *Pew Research Center*, September 2020.

[44] I. Goldberg. *Privacy Enhancing Technologies for the Internet III: Ten Years Later*. Auerbach Publications, 2007.

[45] L. Hanson. Asking for a friend: What if the TikTok algorithm knows me better than I know myself? https://www.gq.com.au/success/opinions/asking-for-a-friend-what-if-the-tiktok-algorithm-knows-me-better-than-i-know-myself/news-story/4eea6d6f23f9ead544c2f773c9a13921, 2021.

[46] K. Hill and S. Mattu. The house that spied on me: The reason I smartened up my house was to find out whether it would betray me. *Gizmodo*, Feb. 2018.

[47] A. L. Holbrook, M. C. Green, and J. A. Krosnick. Telephone versus face-to-face interviewing of national probability samples with long questionnaires: Comparisons of respondent satisficing and social desirability response bias. *Public Opinion Quarterly*, 67(1):79–125, 2003.

[48] G. Horcher. Woman says her Amazon device recorded private conversation, sent it out to random contact. *KIRO 7 News*, May 2018.

[49] Y. Huang, B. Obada-Obieh, and K. Beznosov. Amazon vs. my brother: How users of shared smart speakers perceive and cope with privacy risks. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2020.

[50] Information Commissioner's Office. Project ExplAIn interim report, 2019.

[51] Ipsos. Global public opinion and government use of AI and facial recognition: An Ipsos survey for the World Economic Forum, 2019.

[52] Ipsos. Widespread concern about artificial intelligence, 2019.

[53] B. Jayaraman and D. Evans. Evaluating differentially private machine learning in practice. In *USENIX Security Symposium*, 2019.

[54] A. Jobin, M. Ienca, and E. Vayena. The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1:389–399, September 2019.

[55] R. Kaysen. Is my not-so-smart house watching me? *The New York Times*, Apr. 2018.

[56] P. G. Kelley, Y. Yang, C. Heldreth, C. Moessner, A. Sedley, A. Kramm, D. T. Newman, and A. Woodruff. Exciting, Useful, Worrying, Futuristic: Public perception of artificial intelligence in 8 countries. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21)*, page 627–637, 2021.

[57] P. G. Kelley, Y. Yang, C. Heldreth, C. Moessner, A. M. Sedley, and A. Woodruff. "Mixture of amazement at the potential of this technology and concern about possible pitfalls": Public sentiment towards AI in 15 countries. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 44(4):28–46, 2021.

[58] C. F. Kerry. Protecting privacy in an AI-driven world. *Brookings*, Feb. 2020.

[59] A. Kozyreva, P. Lorenz-Spreen, R. Hertwig, S. Lewandowsky, and S. M. Herzog. Public attitudes towards algorithmic personalization and use of personal data online: Evidence from Germany, Great Britain, and the United States. *Humanities and Social Sciences Communications*, 8:1–11, 2021.

[60] T. Ø. Kuldova. Imposter paranoia in the age of intelligent surveillance: Policing outlaws, borders and undercover agents. *Journal of Extreme Anthropology*, 4(1):45–73, 2020.

[61] J. Lau, B. Zimmerman, and F. Schaub. Alexa, are you listening? Privacy perceptions, concerns and privacy-seeking behaviors with smart speakers. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW), Nov. 2018.

[62] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60, 2020.

[63] Lloyd's Register Foundation. World Risk Poll Report 2019, 2020.

[64] S. Maheshwari. Hey, Alexa, what can you hear? And what will you do with it? *The New York Times*, Mar. 2018.

[65] N. Malkin, J. Deatrick, A. Tong, P. Wijesekera, S. Egelman, and D. Wagner. Privacy attitudes of smart speaker users. *Proceedings on Privacy Enhancing Technologies*, 2019(4):250–271, 2019.

[66] K. Manheim and L. Kaplan. Artificial intelligence: Risks to privacy & democracy. *Yale Journal of Law and Technology*, 106, 2010.

[67] J. Markman. Massive IoT hacks should lead to positive change. *Forbes*, Oct. 2016.

[68] D. McCauley. What the internet of things means for consumer privacy. *The Economist Intelligence Unit Limited*, 2018.

[69] N. McDonald, S. Schoenebeck, and A. Forte. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for CSCW and HCI practice. In *Proceedings of the 22nd ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW 2019)*, 2019.

[70] C. Mele. Bid for access to Amazon Echo audio in murder case raises privacy concerns. *The New York Times*, Dec. 2016.

[71] N. Meng, D. Keküllüoğlu, and K. Vaniea. Owning and sharing: Privacy perceptions of smart speaker users. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–29, 2021.

[72] T. Moynihan. Alexa and Google Home record what you say. but what happens to that data? *Wired*, Dec. 2016.

[73] Mozilla. We asked people around the world how they feel about artificial intelligence. Here's what we learned., 2019.

[74] L.-M. Neudert, A. Knuutila, and P. N. Howard. Global attitudes towards AI, machine learning & automated decision making: Implications for involving artificial intelligence in public service and good governance. *Oxford Internet Institute*, 2020.

[75] Northeastern University and Gallup. Optimism and anxiety: Views on the impact of artificial intelligence and higher education's response, January 2018.

[76] N. Nthala and E. Rader. Towards a conceptual model for provoking privacy speculation. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, page 1–8, New York, NY, 2020. Association for Computing Machinery.

[77] M. Oswald. Artificial intelligence (AI) & explainability citizens' juries report, 2019.

[78] L. Ouchchy, A. Coin, and V. Dubljević. AI in the headlines: The portrayal of the ethical issues of artificial intelligence in the media. *AI & Society*, 35:927–936, 2020.

[79] N. Papernot, M. Abadi, U. Erlingsson, I. Goodfellow, and K. Talwar. Semi-supervised knowledge transfer for deep learning from private training data. *arXiv preprint arXiv:1610.05755*, 2016.

[80] A. Pustozerova and R. Mayer. Information leaks in federated learning. In *Proceedings of the Network and Distributed System Security Symposium*, volume 10, 2020.

[81] E. Rader and R. Gray. Understanding user beliefs about algorithmic curation in the Facebook news feed. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 173–182, 2015.

[82] L. Rainie, C. Funk, M. Anderson, and A. Tyson. AI and human enhancement: Americans' openness is tempered by a range of concerns. *Pew Research Center*, March 2022.

[83] S. Ruoti, J. Andersen, D. Zappala, and K. Seamons. Why Johnny still, still can't encrypt: Evaluating the usability of a modern PGP client. *arXiv preprint arXiv:1510.08555*, 2015.

[84] A. Russakovskii. Google is permanently nerfing all Home Minis because mine spied on everything I said 24/7. *Android Police*, Oct. 2017.

[85] N. J. Salkind, editor. *Encyclopedia of Research Design*, volume 1. Sage, 2010.

[86] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 1528–1540, 2016.

[87] F. M. Shipman and C. C. Marshall. Ownership, privacy, and control in the wake of Cambridge Analytica: The relationship between attitudes and awareness. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, page 1–12, 2020.

[88] A. Singh. Democratising decisions about technology: A toolkit. Technical report, The royal society for arts, manufactures, and commerce (RSA), 2019.

[89] A. Smith. Public attitudes toward computer algorithms. *Pew Research Center*, Nov 2018.

[90] J. Swart. Experiencing algorithms: How young people understand, feel about, and engage with algorithmic news selection on social media. *Social Media + Society*, 7(2), 2021.

[91] O. Tene and J. Polonetsky. A theory of creepy: technology, privacy and shifting social norms. *Yale Journal of Law & Technology*, 16:59, 2013.

[92] The European Commission. Special Eurobarometer 460: Attitudes towards the impact of digitisation and automation on daily life, May 2017.

[93] The Jefferson Center. The citizens' jury handbook, 2004.

[94] B. Ur, P. G. Leon, L. F. Cranor, R. Shay, and Y. Wang. Smart, useful, scary, creepy: Perceptions of online behavioral advertising. In *Proceedings of the Eighth Symposium on Usable Privacy and Security (SOUPS '12)*, pages 1–15, 2012.

[95] J. Warshaw, N. Taft, and A. Woodruff. Intuitions, analytics, and killing ants: Inference literacy of high school-educated adults in the US. In *Proceedings of the Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*, pages 271–285, 2016.

[96] D. M. West. Brookings survey finds divided views on artificial intelligence for warfare, but support rises if adversaries are developing it. *Brookings*, August 2018.

[97] R. Willis. The use of composite narratives to present interview findings. *Qualitative Research*, 19(4):471–480, 2019.

[98] M. L. Wilson, E. H. Chi, S. Reeves, and D. Coyle. RepliCHI: The Workshop II. In *CHI Extended Abstracts '14*, page 33–36, 2014.

[99] C. Wood. Devices sprout ears: What do Alexa and Siri mean for privacy? *Christian Science Monitor*, Jan. 2017.

[100] A. Woodruff, Y. A. Anderson, K. J. Armstrong, M. Gkiza, J. Jennings, C. Moessner, F. Viegas, M. Wattenberg, L. Webb, F. Wrede, and P. G. Kelley. "A cold, technical decision-maker": Can AI provide explainability, negotiability, and humanity? *arXiv preprint arXiv:2012.00874*, 2020.

[101] J. Wright, D. Leslie, C. Raab, F. Ostmann, and M. B. Briggs. Privacy, agency and trust in human-AI ecosystems: Interim report. *The Alan Turing Institute*, 2021.

[102] Y. Yang and N. Liu. China survey shows high concern over facial recognition abuse. *Financial Times*, Dec. 2019.

[103] YouGov. International technology report 2021: Automation & AI, 2021.

[104] B. Zhang and A. Dafoe. Artificial intelligence: American attitudes and trends. *Available at SSRN 3312874*, 2019.

[105] B. Zhang, N. Wang, and H. Jin. Privacy concerns in online recommender systems: influences of control and user data input. In *10th Symposium On Usable Privacy and Security (SOUPS 2014)*, pages 159–173, 2014.

[106] S. Zheng, N. Apthorpe, M. Chetty, and N. Feamster. User perceptions of smart home IoT privacy. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW), 2018.

Appendix

7 Survey Instrument – Select items

Exposure to AI
In the past 12 months, how much have you heard about Artificial Intelligence (AI)?

- A great amount
- A lot
- A moderate amount
- A little bit
- Nothing at all

Knowledge – best description
In your own understanding, which of the following best describes Artificial Intelligence (AI)?

- Any advanced technology
- A system of connected devices
- Technology that can learn or think
- Robot
- Fake News
- Not sure

Knowledge – best example
Which of the following do you think is the best example of Artificial Intelligence (AI)?

- Self-driving car
- Computer
- Fast internet connection
- Spreadsheet

Next 10 Years – privacy
In the next ten years, what do you think will happen in [COUNTRY LABEL] because of Artificial Intelligence (AI)?

- More Privacy
- No Change
- Less Privacy
- Don't know

For each of the questions in Table 1 a parallel Next 10 Years question was asked, e.g., "More jobs created" vs. "More jobs lost" or "Better healthcare" vs. "Worse healthcare", etc.

How will AI affect privacy?
Now we would like to ask you to think about Artificial Intelligence (AI) and privacy. In what ways will Artificial Intelligence (AI) affect privacy in the future? Please be specific.

{Open-end}

8 More Privacy

While the preponderance of our respondents expect AI to have a negative effect on privacy, some expect it to have a positive effect. While many respondents did not offer an explanation, some shared reasons they are optimistic that AI will protect people's privacy, and often connected these ideas with security protection as well.

While one might expect that positive responses may have just been a "halo effect" due to a general belief that AI will affect everything in society in a positive way, some respondents shared specific reasons why they believed AI could truly improve their privacy. We detail four of those types of responses here:

AI defends against hackers. Some respondents suggested AI will keep people's information more secure by defending it against malicious actors. Some even suggested AI could use its learning capabilities to profile and defeat hackers. Others proposed that AI would be useful in detecting security or data breaches.

> AI can identify scams and keep private information safe –*Australia*

> AI can prevent you from being hacked and information is therefore more secure. –*Germany*

> Reinforcing defense capabilities by learning attack patterns against hacking –*South Korea*

> If used properly it could continuously monitor your personal data and search for breaches –*United States*

AI provides safe storage. Some respondents also believe AI keeps data safe. AI was associated with safe storage, secure networks, and encryption. For example, respondents observed that in the era of AI, records will be digitized rather than remaining in paper format and will be securely stored and encrypted.

> Massive data will be stored on the AI side, paper files will be eliminated, and privacy will be effectively protected –*China*

> There will be less reliance on humans to protect the privacy of users, and AI will be able to create more complex systems to encrypt and protect our data –*Australia*

> AI will likely improve privacy since the information fed into the computer is safely stored and no person handles it directly –*Kenya*

AI provides advanced authentication. Some respondents mentioned that AI's authentication capabilities offers them

improved privacy or security as compared with alphanumeric passwords. For example, they called out the use of AI-driven biometric affordances such as facial or fingerprint recognition to unlock phones or voice recognition to authenticate a user to a robot or assistant.

> Artificial intelligence may increase privacy in general if someone is able to set their devices to recognize only them but not strangers. –*Kenya*

> I don't know for sure, but you can have a fingerprint recognition system without password that brings security against hackers. –*Brazil*

AI reduces human involvement. Some participants shared that AI will improve privacy by reducing human-human interaction and human involvement in data processing tasks, and suggested that AI can handle information more reliably and discreetly than humans.

> I imagine AI will allow people to interface with intelligent computers rather than people for sensitive matters like banking. It might make things more secure. –*United States*

> less contact with human hands would mean there would be minimal loss of data or selling of data. –*Kenya*

> If artificial intelligence manages privacy, leakage by humans will be eliminated. –*Japan*

In public opinion polling, the position that AI may have a positive effect on privacy is often dismissed as naive or non-specific. For example, positive responses may reflect a general belief that AI will affect everything in a very positive way. Notably, however, some respondents shared specific reasons that would likely be viewed as legitimate by privacy and security experts.

Country		**Australia**	**Germany**	**Japan**	**South Korea**	**United States**
HDI Rank		5th	9th	19th	19th	21st
Languages offered		English	German	Japanese	Korean	English
Weighting		age, gender, education, region, smartphone OS	age, gender, education, region, smartphone OS	age, gender, education, region, smartphone OS	age, gender, education, region, smartphone OS	age, gender, education, region, smartphone OS, race, HH income, metropolitan status
Respondents		1000	1001	1001	1000	1002
Gender		47% men 52% women	49% men 50% women	53% men 47% women	46% men 53% women	51% men 48% women
Age	16-24	11%	15%	10%	16%	10%
	25-34	15%	16%	23%	19%	13%
	35-44	17%	16%	22%	23%	15%
	45-54	11%	14%	15%	27%	16%
	55-64	17%	15%	11%	11%	22%
	65+	30%	23%	18%	3%	24%
Education	Some primary/secondary	21%	32%	4%	10%	12%
	Completed high school	16%	17%	29%	15%	25%
	Some college or vocational	29%	27%	17%	6%	28%
	Completed Bach. or more	33%	22%	50%	66%	35%

Country		**Russia**	**China**	**Brazil**	**Philippines**	**Kenya**
HDI Rank		52nd	79th	87th	116th	152nd
Languages offered		Russian	Chinese	Brazilian Portuguese	English, Tagalog	English
Weighting		age, gender, education, region, smartphone OS	age, gender, education, region, smartphone OS	age, gender, education, region, smartphone OS	age, gender, education, region, smartphone OS	age, gender, education, region, smartphone OS
Respondents		1000	1004	1001	1000	1002
Gender		46% men 54% women	54% men 46% women	46% men 54% women	45% men 55% women	52% men 48% women
Age	16-24	10%	16%	34%	30%	26%
	25-34	35%	26%	23%	35%	39%
	35-44	20%	25%	22%	21%	20%
	45-54	21%	23%	12%	10%	10%
	55-64	12%	8%	8%	3%	4%
	65+	3%	2%	2%	1%	1%
Education	Some primary/secondary	3%	3%	12%	3%	4%
	Completed high school	6%	10%	29%	15%	9%
	Some college or vocational	27%	20%	17%	29%	38%
	Completed Bach. or more	63%	66%	42%	53%	49%

Table 3: Country details, respondent summary and demographics. Percentages for gender and education may not add up to 100% due to participants who preferred to self-describe or not disclose. All numbers and percentages here are unweighted; throughout the rest of the paper all numbers are weighted, by country, based on the weighting variables above. We show HDI ranks from the 2022 Human Development Report https://hdr.undp.org/content/human-development-report-2021-22, which uses HDI values from 2021, aligning with the dates of our survey deployment.

AI Knowledge – best description	**Global average**	**Australia**	**Brazil**	**China**	**Germany**	**Japan**	**Kenya**	**Philippines**	**Russia**	**South Korea**	**United States**
Technology that can learn or think	47%	49%	50%	38%	53%	53%	38%	41%	57%	27%	63%
Robot	21%	18%	17%	24%	20%	18%	27%	20%	23%	30%	10%
Any advanced technology	15%	13%	12%	14%	5%	21%	19%	24%	6%	24%	9%
A system of connected devices	10%	8%	14%	20%	10%	6%	7%	8%	6%	14%	3%
Fake news	1%	3%	1%	1%	2%	1%	1%	1%	1%	2%	1%
Not sure	7%	9%	5%	3%	10%	2%	10%	5%	7%	3%	13%
Refused	0%	–	–	–	–	–	–	–	–	–	1%
AI Knowledge – best example											
Self-driving car	57%	68%	44%	58%	61%	63%	39%	55%	59%	55%	71%
Computer	27%	24%	34%	13%	23%	24%	45%	32%	30%	23%	22%
Fast internet connection	13%	6%	21%	25%	12%	11%	14%	10%	9%	18%	3%
Spreadsheet	3%	3%	1%	4%	5%	2%	2%	3%	3%	4%	1%
Refused	0%	–	–	–	–	–	–	–	–	–	3%

Table 4: Summary results for our two AI knowledge questions. Item selection was based on open-ended responses in our own previous research as well as iterative piloting in an online survey platform.

Global Averages		*Codes*
	Privacy Themes	
31%	**Highly Personal**	Personal Data, Intrusive, Tracking, Corporations/Companies, Location, Personalization, Data Analysis, Targeting, Social Media, Other Services and Retail, Ads, Daily life, Ads Follow Me, Self-driving, Other Internet Services, Profit, Home Appliances, Search Engine, Deep Fakes, Amazon, Google [the company],...
29%	**Data at Risk**	Hacking, Collection, Available, Needs Data, Connected, Bad Purposes, Big Data, Devices, Phone, Internet, Collation,...
12%	**State and Surveillance**	Listening, Surveillance, Governments, Biometrics, Security Cameras, Camera, Control Population, Conversation, Criminals, Catch Criminals, Country, Assistant,...
5%	**Without Consent**	Consent, Unaware.
	Privacy Sentiment	
58%	Negative sentiment or part of a theme	Codes used in all themes above, and: Less Privacy, No Privacy, Facilitation, Danger, Fear, Privacy, Hurt.
12%	Positive sentiment	More Privacy, Security, Less Human Contact.
	Overall Response Quality	
75%	Expressed any privacy statement	All codes above, and other non-sentiment privacy codes, including: No Effect on Privacy, Data, Effect on Privacy Depends, It Will Impact Privacy, Other Remediation, Other Privacy, Regulation, Too Early to Tell,...
18%	Don't know	I don't know, Inarticulate, Blank or no comment, Unable to code.
8%	Any other, unrelated response	Any other code, including: Technology, Computer, Advanced, Inevitable, Useful, AI Takes Over, Other, Job loss, Productivity, Learn, Future, Think, Robot, Helpful, Concern, Machine, Other Applications, AI Replaces Humans, Makes Mistakes, Improves Quality of Life, Program, Could Go Either Way, Home/House, Good and bad, Automated, Intelligence, Benefits, Unfair, Other Sentiment, Responsibility, Powerful, Humans Get Less Skilled, Bad, Autonomy, Not Trustworthy, Assist, Communication, Mechanical,...

Table 5: Open-ended codes used to create each theme, sentiment grouping, and to describe overall response quality. This table shows all codes that had 25 or more uses, totaled across all countries. Codes were assigned to themes based on emergent clustering. For example, the "Corporations/Companies" code was assigned to the "Highly Personal" theme because mentions of corporations/companies in the context of the privacy question were typically about invasive corporate use of personal data.

Factor	Control	Treatment	Odds	P>|z|
Country	United States	Japan	0.24	0.000
Country	United States	Russia	0.40	0.000
Country	United States	China	0.42	0.000
Country	United States	Kenya	0.47	0.000
Country	United States	Philippines	0.48	0.000
Country	United States	South Korea	0.50	0.000
Country	United States	Brazil	0.52	0.000
Country	United States	Australia	0.61	0.000
Country	United States	Germany	0.61	0.000
Age	16-24	25-34	0.89	0.088
Age	16-24	35-44	0.99	0.879
Age	16-24	55-64	1.03	0.763
Age	16-24	45-54	1.11	0.146
Age	16-24	65+	1.53	0.000
Gender	Male	Female	1.07	0.099
Gender	Male	Prefer To Self-Describe	4.31	0.070
Education	Some primary or secondary	Completed high school	1.20	0.017
Education	Some primary or secondary	Some college or vocational studies	1.24	0.003
Education	Some primary or secondary	Completed Bachelor's or more	1.59	0.000
Definition of AI	Not Sure	Any Advanced Technology	1.94	0.000
Definition of AI	Not Sure	Robot	2.32	0.000
Definition of AI	Not Sure	A System Of Connected Devices	2.39	0.000
Definition of AI	Not Sure	Fake News	2.69	0.000
Definition of AI	Not Sure	Technology That Can Learn Or Think	3.13	0.000
Example of AI	Spreadsheet	Fast Internet Connection	1.28	0.098
Example of AI	Spreadsheet	Computer	1.62	0.001
Example of AI	Spreadsheet	Self-Driving Car	2.27	0.000
Exposure to AI	A Moderate Amount	A Great Amount	0.76	0.000
Exposure to AI	A Moderate Amount	Nothing At All	0.84	0.027
Exposure to AI	A Moderate Amount	A Lot	0.95	0.365
Exposure to AI	A Moderate Amount	A Little Bit	1.23	0.001

Table 6: Odds of a respondent believing they will have "less privacy" in ten years due to AI when holding all factors but one constant. Reporting includes all data, irrespective of $p < 0.05$.

Factor	Control	Treatment	Odds	P>|z|
Country	United States	Russia	0.23	0.000
Country	United States	Germany	0.37	0.000
Country	United States	South Korea	0.53	0.000
Country	United States	China	0.70	0.002
Country	United States	Japan	0.74	0.008
Country	United States	Brazil	0.75	0.011
Country	United States	Australia	1.12	0.285
Country	United States	Philippines	1.14	0.239
Country	United States	Kenya	1.16	0.178
Age	16-24	65+	0.52	0.000
Age	16-24	55-64	0.59	0.000
Age	16-24	35-44	0.63	0.000
Age	16-24	45-54	0.69	0.000
Age	16-24	25-34	0.78	0.000
Gender	Male	Female	1.09	0.088
Education	Some primary or secondary	Completed high school	1.39	0.000
Education	Some primary or secondary	Some college or vocational studies	1.66	0.000
Education	Some primary or secondary	Completed Bachelor's or more	1.95	0.000
Definition of AI	Not Sure	Fake News	1.90	0.026
Definition of AI	Not Sure	Robot	2.59	0.000
Definition of AI	Not Sure	Any Advanced Technology	2.61	0.000
Definition of AI	Not Sure	A System Of Connected Devices	3.36	0.000
Definition of AI	Not Sure	Technology That Can Learn Or Think	3.62	0.000
Example of AI	Spreadsheet	Fast Internet Connection	1.80	0.007
Example of AI	Spreadsheet	Computer	2.51	0.000
Example of AI	Spreadsheet	Self-Driving Car	3.21	0.000
Exposure to AI	A Moderate Amount	Nothing At All	0.71	0.000
Exposure to AI	A Moderate Amount	A Great Amount	0.86	0.049
Exposure to AI	A Moderate Amount	A Little Bit	0.87	0.037
Exposure to AI	A Moderate Amount	A Lot	1.06	0.382

Table 7: Odds of a respondent sharing a theme coded as **Data at Risk** in their top-of-mind concerns related to privacy and AI. Reporting includes all data, irrespective of $p < 0.05$.

Factor	Control	Treatment	Odds	P>|z|
Country	United States	Russia	0.31	0.000
Country	United States	Germany	0.32	0.000
Country	United States	South Korea	0.54	0.000
Country	United States	Japan	0.59	0.000
Country	United States	China	0.65	0.000
Country	United States	Brazil	0.74	0.007
Country	United States	Australia	0.81	0.046
Country	United States	Kenya	0.97	0.807
Country	United States	Philippines	1.18	0.117
Age	16-24	65+	0.66	0.000
Age	16-24	45-54	0.86	0.064
Age	16-24	35-44	0.87	0.066
Age	16-24	25-34	0.90	0.147
Age	16-24	55-64	0.95	0.578
Gender	Male	Female	1.06	0.219
Education	Some primary or secondary	Completed high school	1.58	0.000
Education	Some primary or secondary	Some college or vocational studies	2.08	0.000
Education	Some primary or secondary	Completed Bachelor's or more	2.18	0.000
Definition of AI	Not Sure	Fake News	1.45	0.165
Definition of AI	Not Sure	Any Advanced Technology	1.62	0.001
Definition of AI	Not Sure	Robot	1.67	0.000
Definition of AI	Not Sure	A System Of Connected Devices	2.08	0.000
Definition of AI	Not Sure	Technology That Can Learn Or Think	2.59	0.000
Example of AI	Spreadsheet	Fast Internet Connection	1.43	0.067
Example of AI	Spreadsheet	Computer	1.72	0.004
Example of AI	Spreadsheet	Self-Driving Car	2.38	0.000
Exposure to AI	A Moderate Amount	Nothing At All	0.52	0.000
Exposure to AI	A Moderate Amount	A Great Amount	0.74	0.000
Exposure to AI	A Moderate Amount	A Little Bit	0.86	0.025
Exposure to AI	A Moderate Amount	A Lot	0.88	0.039

Table 8: Odds of a respondent sharing a theme coded as **Highly Personal** in their top-of-mind concerns related to privacy and AI. Reporting includes all data, irrespective of $p < 0.05$.

Factor	Control	Treatment	Odds	P>|z|
Country	United States	Russia	0.15	0.000
Country	United States	Germany	0.28	0.000
Country	United States	South Korea	0.36	0.000
Country	United States	Brazil	0.60	0.018
Country	United States	China	0.60	0.016
Country	United States	Kenya	0.73	0.113
Country	United States	Australia	0.81	0.273
Country	United States	Japan	0.88	0.535
Country	United States	Philippines	0.92	0.684
Age	16-24	65+	0.66	0.065
Age	16-24	45-54	0.85	0.324
Age	16-24	35-44	0.91	0.531
Age	16-24	25-34	1.02	0.868
Age	16-24	55-64	1.13	0.500
Gender	Male	Female	1.13	0.219
Education	Some primary or secondary	Completed high school	0.93	0.724
Education	Some primary or secondary	Completed Bachelor's or more	1.42	0.063
Education	Some primary or secondary	Some college or vocational studies	1.47	0.040
Definition of AI	Not Sure	Fake News	0.46	0.406
Definition of AI	Not Sure	A System Of Connected Devices	1.33	0.417
Definition of AI	Not Sure	Any Advanced Technology	1.35	0.362
Definition of AI	Not Sure	Robot	1.42	0.282
Definition of AI	Not Sure	Technology That Can Learn Or Think	2.03	0.023
Example of AI	Spreadsheet	Computer	1.73	0.298
Example of AI	Spreadsheet	Fast Internet Connection	2.16	0.151
Example of AI	Spreadsheet	Self-Driving Car	2.90	0.040
Exposure to AI	A Moderate Amount	Nothing At All	0.51	0.002
Exposure to AI	A Moderate Amount	A Great Amount	0.83	0.231
Exposure to AI	A Moderate Amount	A Little Bit	0.90	0.458
Exposure to AI	A Moderate Amount	A Lot	0.97	0.824

Table 9: Odds of a respondent sharing a theme coded as **Without Consent** in their top-of-mind concerns related to privacy and AI. Reporting includes all data, irrespective of $p < 0.05$.

Factor	Control	Treatment	Odds	P>|z|
Country	United States	Japan	0.24	0.000
Country	United States	China	0.31	0.000
Country	United States	Russia	0.35	0.000
Country	United States	South Korea	0.50	0.001
Country	United States	Philippines	0.56	0.007
Country	United States	Germany	0.58	0.008
Country	United States	Brazil	0.69	0.065
Country	United States	Kenya	1.08	0.699
Country	United States	Australia	1.19	0.319
Age	16-24	35-44	1.39	0.040
Age	16-24	25-34	1.44	0.015
Age	16-24	65+	1.56	0.025
Age	16-24	55-64	1.59	0.015
Age	16-24	45-54	1.62	0.003
Gender	Male	Female	0.70	0.000
Education	Some primary or secondary	Completed Bachelor's or more	1.48	0.031
Education	Some primary or secondary	Some college or vocational studies	1.57	0.013
Education	Some primary or secondary	Completed high school	1.65	0.008
Definition of AI	Not Sure	Fake News	1.56	0.509
Definition of AI	Not Sure	Robot	2.29	0.026
Definition of AI	Not Sure	A System Of Connected Devices	2.80	0.008
Definition of AI	Not Sure	Any Advanced Technology	3.18	0.002
Definition of AI	Not Sure	Technology That Can Learn Or Think	3.18	0.001
Example of AI	Spreadsheet	Fast Internet Connection	2.08	0.155
Example of AI	Spreadsheet	Computer	2.59	0.057
Example of AI	Spreadsheet	Self-Driving Car	2.90	0.031
Exposure to AI	A Moderate Amount	Nothing At All	0.69	0.057
Exposure to AI	A Moderate Amount	A Little Bit	1.00	0.994
Exposure to AI	A Moderate Amount	A Lot	1.07	0.610
Exposure to AI	A Moderate Amount	A Great Amount	1.38	0.024

Table 10: Odds of a respondent sharing a theme coded as **State and Surveillance** in their top-of-mind concerns related to privacy and AI. Reporting includes all data, irrespective of $p < 0.05$.

Investigating Security Indicators for Hyperlinking Within the Metaverse

Maximiliane Windl[1,2], Anna Scheidle[1], Ceenu George[3,4], Sven Mayer[1,2]
[1] *LMU Munich, Germany*
[2] *Munich Center for Machine Learning (MCML), Germany*
[3] *University of Augsburg, Germany*
[4] *TU Berlin, Germany*

Abstract

Security indicators, such as the padlock icon indicating SSL encryption in browsers, are established mechanisms to convey secure connections. Currently, such indicators mainly exist for browsers and mobile environments. With the rise of the metaverse, we investigate how to mark secure transitions between applications in virtual reality to so-called sub-metaverses. For this, we first conducted in-depth interviews with domain experts (N=8) to understand the general design dimensions for security indicators in virtual reality (VR). Using these insights and considering additional design constraints, we implemented the five most promising indicators and evaluated them in a user study (N=25). While the visual blinking indicator placed in the periphery performed best regarding accuracy and task completion time, participants subjectively preferred the static visual indicator above the portal. Moreover, the latter received high scores regarding understandability while still being rated low regarding intrusiveness and disturbance. Our findings contribute to a more secure and enjoyable metaverse experience.

1 Introduction

At the latest, when Facebook renamed itself to Meta and put most of its research efforts into creating an immersive virtual world, the notion of the "metaverse" attracted the public's attention. While employing different approaches, other companies also focus on creating such "shared, open, and perpetual virtual worlds" [32]. For example, Microsoft is creating a collaborative, mixed-reality experience mainly for meetings[1]; Niantic is developing outdoor-capable AR glasses aiming to enrich the real world instead of cutting people out of it[2], and EPIC games and LEGO are developing a secure metaverse experience for children[3]. While many develop on independent applications, the idea of the metaverse links all single applications to one big network, much like the world-wide-web with hyperlinks to transition between them. Therefore, users will transition between different metaverses via hyperlinking frequently and consequently take their identity to unknown environments. Thus, it is only a matter of time before known security risks from browser and mobile environments become relevant threats [23]. As such, the same need as in the world-wide-web will occur – marking and ensuring safe transitions to a new service before revealing one's identity to the new provider. Moreover, users will frequently need to decide whether to consciously enter applications with unknown origins [39]. Hence, we require effective security indicators within the metaverse.

Prior research has shown that security indicators can effectively signal secure transitions to users. An established example represents the padlock icon displayed in browsers next to the URL to signal SSL encryption, cf. [48]. So far, research has primarily focused on security indicators in browsers [11, 24, 33] and mobile environments [35, 52]. Here, researchers investigated the effectiveness of using, for example, icons [25, 45], color coding [45], blinking animations [30], or security images that should create a secret between the user and the application [30]. Going from 2D to 3D space offers many novel ways to represent security indicators, such as size and location. Therefore, we argue that the next step will be to extrapolate from 2D indicators and develop indicators suitable for the metaverse.

This paper investigates security indicators for hyperlinking within the metaverse. For this, we first employed a participatory design approach by conducting in-depth interviews with domain experts (N=8) to understand the general design dimensions for security indicators in VR. Based on the expert interviews' findings, we developed and evaluated the five most promising security indicators (one haptic, one audio, and

[1] https://www.microsoft.com/en-us/mesh
[2] https://nianticlabs.com/news/real-world-metaverse
[3] https://www.epicgames.com/site/en-US/news/the-lego-group-and-epic-games-team-up-to-build-a-place-for-kids-to-play-in-the-metaverse

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.
USENIX Symposium on Usable Privacy and Security (SOUPS) 2023.
August 6–8, 2023, Anaheim, CA, United States.

three visual indicators) for their usability and effectiveness in a user study (N=25). We found that while the five indicators performed equally well regarding pragmatic and hedonic quality, there were considerable differences regarding understandability and disturbance. Moreover, the visual blinking indicator significantly improved the accuracy and speed of understanding secure transitions in VR.

Our contribution is twofold. First, this paper is the first to construct a design space for security indicators in VR. This design space will help researchers and developers create security indicators for VR. Second, our study showed that while the visual blinking indicator placed in the periphery performed best regarding the objective measures accuracy and task completion time, participants subjectively preferred the static visual indicator placed above the portal. Moreover, it received high scores for understandability while still being rated low regarding intrusiveness and disturbance. Our findings have implications for designing metaverse environments by ensuring a more secure and enjoyable VR experience.

2 Related Work

We first present definitions of the term metaverse and research trends. Then, we discuss prior research on security indicators on the web, mobile environments, and mixed reality. Finally, we derive our research questions.

2.1 Metaverse

The term metaverse appeared for the first time in a novel by Stephenson [46], where it is described as a parallel universe where people interact through avatars. While the metaverse attracted attention in research, there exists no common definition. While Park and Kim [37] state that the metaverse does not necessarily use VR and AR technologies, Green and Works [19] found that the metaverse is commonly described as a virtual world that uses VR technology by researching the term's definition across social media, the news, and in the ACM. Moreover, Lee et al. [32] define the metaverse as "a virtual environment blending physical and digital," and Ning et al. [36] add the "interaction of humans and a computer-mediated virtual platform" as other key aspects. *Consequently, we define the metaverse as a connected social environment that uses VR technology in this paper's context.*

There is also research on how a widely adopted metaverse might influence the world. Duan et al. [10] outline how the metaverse can be used for social good. They, for example, describe how a metaverse can improve accessibility by hosting social events so that no travel is required, improve diversity as a metaverse would make it easier to cater to individual needs, and how the metaverse can help humanity as historical landmarks can be rebuilt in VR. However, there is also research on the possible threats of the metaverse. Rosenberg [40], for example, outlines three fundamental risks: 1) The current ubiquitous monitoring of users will get even worse in the metaverse, as a multitude of new features can be tracked, such as where users go, looks at, what they grab, or their vital signs; 2) Manipulation of users might also worsen as it will become hard to differ advertisement from real content, as advertisements might be hidden as simulated people or products; and 3) monetization of users in the metaverse will become an issue as people pay with their data. To counteract these possible negative effects, Rosenberg [40] suggest non-regulatory and regulatory approaches, such as restricting the monitoring and emotional analysis of metaverse users or restricting virtual product placements. Especially the first point, the increased monitoring of users, might lead to privacy issues, as massive amounts of personal data are collected and stored. Indeed, a large stack of research solely focuses on the privacy and security implications of the metaverse [5, 8, 12, 50]. In terms of privacy, key concerns include but are not limited to the extensive amounts of personal data collected to build a digital copy of the real world [5, 12, 50], social engineering hacking [5, 8, 50], online harassment [12], and more specifically spying and stalking [8]. In terms of security, Di Pietro and Cresci [8] predict issues regarding authentication as it might become hard to distinguish humans from machines and issues regarding polarization and radicalization as a uniform, massive metaverse replaces the present plurality of the web. In addition, Wang et al. [50] raise concerns about data tempering attacks that might happen during data communication among various sub-metaverses and privacy leakage that might happen as large amounts of data are transferred. *As privacy and security are significant concerns about the metaverse, especially when transmitting large amounts of private data and transitioning between so-called sub-metaverses, we see a need for researching adequate mitigation measures.*

2.2 Security Indicators

Security indicators alert the user of potential risks or validate the identity of a website or application [30, 47]. Prior research has investigated the effectiveness of security indicators on the web. The padlock icon next to the URL is one of the browser's most widely adopted security indicators, demonstrating an authenticated connection [47]. Whalen and Inkpen [51] found while the padlock icon was mostly recognized, users did not use its interaction functions, and von Zezschwitz et al. [49] found that many users still misunderstand the icon. While it only indicates connection security, many people misattribute general privacy, security, and trustworthiness to it [49]. Lee et al. [30] tested the effectiveness of security images during login. A security image is supposed to prevent phishing attacks by displaying a personalized image and caption. Yet, researchers found that most users still log in, even if the image is missing [30, 43]. However, users' attention to security images can be improved by adding a visual effect, such as a blinking animation [30] and making them interactive, such as

requiring the user to find and click the image [21].

Prior research investigated security indicators for mobile devices. For example, Zhang et al. [53] tested the effectiveness of warning notices that alerted users of untrusted certificates. They found that the warnings increased users' perceived threat to their personal information. Another line of research focuses on informing users about possible privacy and security threats before installing an application by providing information in the app store. Choe et al. [6] compared positively and negatively framed visuals and found that they influenced users' app installation decisions. Rajivan and Camp [38] explored the usage of icons in the app store to provide information about applications that access private data. Icons positively influenced the app ratings and increased users' subjective perceptions of the app's privacy and security [38]. However, the most prominent of these indicators is the "privacy nutrition label" [26, 27]. Although such labels are currently deployed in both major app stores[4, 5], they have experienced criticism as they are not prominently placed, use confusing terminology, and are inconsistent with the apps' privacy policies [7].

Previous research on mixed reality security indicators almost solely focuses on indicators for secure authentication by developing techniques to shield users' input from external observers [1, 14, 15, 16, 17]. For this, researchers used randomly color-coded visual cues [1], 3D objects [16, 17], and spatial and virtual targets [14]. In the augmented reality (AR) context, prior research anticipates that future AR systems will run multiple applications simultaneously to share input and output devices, exposing data and APIs to each other [39]. This entails risks like clickjacking attacks that trick users into clicking on malicious interface elements [39]. Moreover, users need to know the origin of content to judge if it is trustworthy, especially when sensitive data is shared across applications [39]. Recognizing these dangers, Hosfelt et al. [22] developed different concepts for security indicators for transitions in the immersive web: A logo, a sigil, and a customizable agent. At the time of this paper, it is the only prior work focusing specifically on security indicators for VR transitions. While most participants preferred the agent, the logo performed best regarding the error rate mainly because participants forgot what their sigil or agent looked like. *While signaling secure origins and transitions have been recognized as important in prior research, security indicators for VR have been scarcely researched so far. Hence, we require more research on how to implement indicators that verify the origin and security status of applications and contents in VR.*

2.3 Summary and Research Questions

Prior research raised significant privacy and security concerns regarding the metaverse [5, 8, 12, 50]. Especially as large amounts of data will be transmitted between sub-metaverses [50], users need indicators to verify the origins of content [39]. Yet, before implementing indicators, we first need to know what they can look like. Therefore, we pose our first research question (**RQ1**): *What are the general design dimensions for security indicators in the metaverse?* Researchers found that while security indicators can be effective measures to indicate a secure connection or origin, several have shortcomings preventing them from fulfilling their goals [7, 30, 43]. Hence, we ask our second research question (**RQ2**): *Which security indicators are the most effective?* Yet, for users to willingly use indicators, they must be usable. Thus, our third research question (**RQ3**) is: *Which security indicators are the most usable?* Security indicators must be noticeable and understandable to fulfill their goal while not being intrusive or disturbing. Hence, our fourth research question (**RQ4**) is: *Which security indicators have the best notification qualities?* Lastly, weighing all these qualities against each other, we investigate the best tradeoff with our last research question (**RQ5**): *Which security indicators would participants like to use in their daily VR experience?*

3 A Design Space for VR Security Indicators

As research on security indicators in the metaverse is scarce, we conducted eight semi-structured expert interviews from industry and academia to understand their general design dimensions (**RQ1**). Interviews allowed us to follow up on the experts' ideas and start an in-depth discussion on them.

3.1 Procedure

Before we started the interview, we provided experts with an informed consent form and practical information, such as the session duration and confidentiality. We then started with introductory questions about our experts' general familiarity and experience with security indicators, followed by their familiarity with VR. After that, we introduced the security issues that might arise in the metaverse when transitioning between applications and the interviews' goal of understanding the general design dimensions of security indicators for usage in VR. To spark our experts' creativity, we presented approaches that prior work had found to be effective for security indicators and to attract users' attention in VR. This included different placements, i.e., in the periphery [20, 33] or the user's focus area [20], the different forms of representation and customization, such as 2D or 3D objects or customized avatars, and the different ways to draw user attention, such as using a pulsating [31] or blinking [20] effect. We then asked our experts to envision security indicators they consider suitable to signal secure transitions in VR, whereby we advised them to describe the different parameters, form factors, and functionalities in detail.

[4] https://www.apple.com/privacy/labels/
[5] https://blog.google/products/google-play/data-safety/

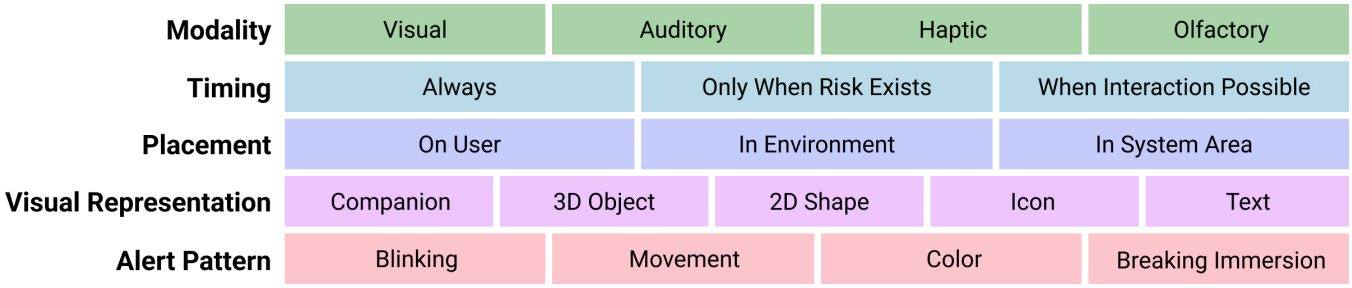


Figure 1: A design space for security indicators in VR based on expert interviews.

3.2 Participants

Eight experts took part in the interviews. We recruited our experts through our personal network, followed by snowball sampling. To qualify as an expert, the participant had to have at least 3 years of experience with VR, usable security, and/or privacy, see Table 1. All participants had either an IT or usability background. In addition, all experts stated that they have experience with security indicators in their daily life or work, and three experts already had experience with security indicators in VR. We did not compensate the experts.

3.3 Results

We transcribed 257.25 minutes ($M = 32.16$, $SD = 6.64$) of audio material, which we recorded during the eight interviews and analyzed the data using thematic analysis [4] and Atlas.ti. More precisely, we followed the *theoretical approach* as outlined in Braun and Clarke [4], where one "code[s] for a specific research question." For that, two researchers first coded all transcribed interviews, after which a third researcher joined to create the code groups and themes in multiple hour-long sessions. This process resulted in two themes: DESIGN SPACE and CONTEXT SPACE.

3.3.1 Design Space

We extracted five themes that describe the dimensions of security indicators in VR, which we used to create a design space, see Figure 1. These five themes are MODALITY, TIMING, PLACEMENT, VISUAL REPRESENTATION, and ALERT PATTERN.

Modality. Our experts discussed four general modalities that can be used to deliver the security indicator. All experts suggested at least one visual security indicator, especially because of its simplicity: *"I would probably go for something simple. For this, a visual cue is actually quite good."* Next to this, our experts also suggested *Auditory* (E1, E3, E6, E8), *Haptic* (E1, E3, E4, E6, E8), and *Olfactory* (E3) indicators.

Table 1: Demographics of our interviewed experts: Their experience, and whether they work in industry or academia.

ID	Experience	Sector
1	Usable Security, Collaborative VR, Presence	Academia
2	Software Engineering, Mobile, XR	Industry
3	Usable Security, Authentication, XR	Academia
4	Software Engineering, XR	Industry
5	Mobile Security and Privacy	Industry
6	Software Engineering, XR	Industry
7	Usable Security and Privacy, Gaze-based Systems	Academia
8	Usable Security and Eye Tracking	Academia

Timing. Our experts named three different timings suitable to display security indicators. One suggestion was to show the indicator *Always*. While E7 opposed this as they feared it might be *"distracting"* and *"annoying,"* E4 suggested implementing such an indicator subtle but still obtrusive, comparing it to the green light indicating an active camera in laptops (E4). Another suggestion was to display the indicator *Only When Risk Exists* (E3, E4), i.e., when it becomes *"relevant (E4)."* The suggestion made by most experts was to display the indicator *When Interaction [is] Possible* (E2, E3, E6, E8), so only displaying the security indicator when a user is *"close enough to interact (E6)"* with a portal or when users have the *"option to change to another environment (E3)."*

Placement. Our experts discussed three general placements for the security indicators. The option most frequently mentioned was placing the indicator *On [the] User* (E1, E2, E4, E5, E6, E8), for example, directly in the user's field of view: *"Perhaps directly centered in the middle (E5),"* or in the periphery around the user (E6). Apart from that, our experts also suggested placing the indicator *In [the] Environment* near the transition or portal (E2, E5, E6, E7, E8) or in the *System Area* (E4, E5, E7), such as it is done in browsers for the padlock.

Visual Representation. Our experts also discussed five different visual representations. They extensively discussed a more playful variant in the form of a *Companion* that actively

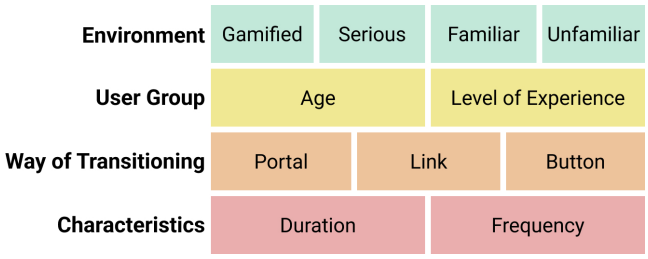


Figure 2: A context space influencing the suitability of different characteristics of security indicators in VR based on expert interviews.

warns the user whenever security issues occur. Such a *Companion* could take different forms, such as an animal (E2), avatar (E5), or even a small robot (E7). However, half of our experts considered *Companions* unsuitable for the security context since they found them either too playful or complex or only understandable with the help of onboarding (E1, E2, E3, E5). While two experts suggested using *3D Objects* (E2, E6), most experts favored simple *2D Shapes* (E3, E4, E5, E6, E8), for example, in the form of a red dot (E4, E3, E7). Other suggestions included porting the padlock *Icon* to VR (E3) or using *Text* (E6). Yet, E6 also discussed the challenges of using *Text* as a security indicator: *"People always click it away and reading in VR is no fun anyway (E6)."*

Alert Pattern. The experts discussed four different alert patterns to draw users' attention to the security indicator. The experts most often suggested to *Break Immersion*, by, for example, either playing an unpleasant sound (E6) or more subtle sounds like a beeping noise (E3) or a whisper (E8). Other suggestions to break immersion included displaying the indicator directly in the user's field of view (E2, E3, E4), dispensing an unpleasant smell (E3), using thermal feedback (E3), or letting the controller vibrate using an obtrusive pattern, such as an elevated heartbeat (E6). Here, a secure transition would be indicated using a calm pattern, and an insecure transition by an elevated heart rate pattern (E6): *"Something exciting that feels somewhat stressful."* Next to this, experts suggested alerting users by changing the *Color* (E3, E4, E5, E6, E8) of the security indicator, using a *Blinking* effect (E2, E4, E5, E6), or using *Movement*, for example, increasing the rotation speed (E2) or changing the size of the indicator (E3).

3.3.2 Context Space

Next to the general design dimensions, our experts also discussed different contextual factors influencing the suitability of the different indicators. We used these insights to create a context space, depicted in Figure 2. It has four levels: ENVIRONMENT, USER GROUP, TYPE OF TRANSITION, and CHARACTERISTICS OF TRANSITION.

Environment. Our experts discussed how the type of environment influences the suitability of the different security indicators. Here, our experts emphasized that the indicator as a companion only fits *Gamified* environments or applications specifically designed for children (E5): *"If it's a game [...] and weird creatures are running around all the time anyway, suddenly some thing jumps around the corner and says: Here, you're going into the wrong world or something. That would be okay."* In contrast, neutral indicators might fit more in *Serious* environments, such as meeting rooms (E5, E6). Additionally, our experts differed between *Familiar* and *Unfamiliar* environments. While unfamiliar environments call for stricter standardization since it might otherwise be hard or impossible for users to differentiate security indicators from other elements, familiar environments allow for more experimental indicators (E8).

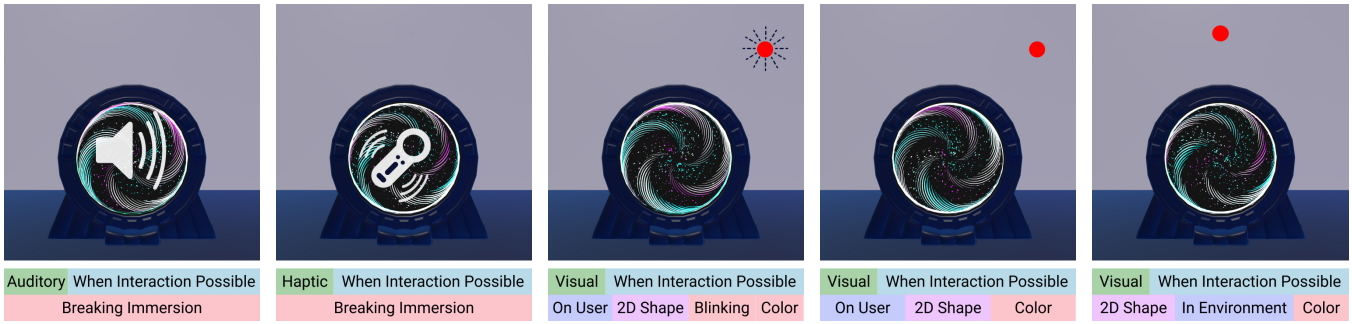
User Group. One factor related to the type of environment mentioned previously is the *Age* of the user. While more serious and neutral indicators are suitable for adults, playful indicators might be used for children: *"I think a stuffed animal [...] would be quite suitable for children (E5)."* The other differentiating factor is the *Level of Experience*. While warning notifications containing text might be suitable to teach inexperienced users the meaning of the indicator, more experienced users might be annoyed by extensive explanations and, thus, prefer more concise indicators (E6).

Way of Transitioning. Transitioning between applications might happen in different ways. While some transitions happen through *Portals*, others, for example, happen through an invitation that includes a *Link* or confirmation *Button* (E2), which in turn determines where a security indicator should and could be placed, as E2 explained: *"The information about whether the transition is safe is not so relevant if I stand ten meters away from the portal. But if I am really close to the portal, it is because only then can I start the transition."*

Characteristics of Transitions. Here, the *Duration* and *Frequency* of a transition matter (E3, E8). Quick transitions call for simple indicators that can be understood quickly, as E3 explains: *"Often, transitions happen very quickly. And then you also have to react very quickly (E3)."* Moreover, while transitions that happen very rarely need to be more obtrusive and contain additional information so that the user understands them, security indicators for frequent transitions can be reduced to simpler versions as the user is already familiar with them (E8).

3.3.3 Indicator Selection

We selected five different indicators to test in our user study, considering our experts' feedback and taking additional design constraints into account. The indicators can be seen in



(a) Audio (b) Haptic (c) Blinking (d) Peripheral (e) Static

Figure 3: The five indicators evaluated in our user study and the design dimensions used to create them.

Figure 3. Our most vital consideration was not restricting the metaverse designers' freedom by placing the indicators directly in the 3D environment. Here, we ensured that the indicator does not occupy more than one sense at a time (e.g., visual and auditory), as this would drastically reduce the expressive freedom of designers and developers of VR environments. Thus, we only use one modality (i.e., one sense) at a time. Moreover, we also wanted to investigate all MODALITIES (see Figure 1) the experts suggested at least once to explore the full design potential (except for olfactory indicators, as dispensing smells is not technically feasible at the moment). In addition, we designed two more visual indicators that, however, in contrast to the other indicators, do not interfere with the 3D environment design as they are not placed in the environment (see Figure 1, *In [the] Environment*) but anchored in the user's field of view (see Figure 1, *On [the] User*). These considerations led to the following five indicators: (1) An *audio* indicator in the form of an unpleasant warning sound (constant 1000Hz beep) as suggested by E3, whereby a secure transition is indicated by no sound similar to a fire or ambulance siren that only sounds when there is danger. (2) A *haptic* indicator using the heart rate pattern suggested by E6, whereby a calm heart rate pattern indicates a secure transition, and an elevated pattern an insecure transition. The following three indicators were color-coded 2D shapes, whereby secure transitions are colored green and have square shapes, and insecure transitions are red and round (we added the shapes to support acceptability needs). These indicators differ in their placement and whether they used a blinking effect: (3) a *visual peripheral blinking* indicator, (4) A *visual peripheral static* indicator, and (5) a *visual static* indicator above the portal.

4 Indicator Evaluation

We conducted a lab study with 25 participants to evaluate the security indicators for their effectiveness (**RQ2**), usability (**RQ3**), notification qualities (**RQ4**), and overall user preference (**RQ5**). We developed a maze VR game containing several portals which participants could use to teleport themselves closer to the exit. We used a maze to (1) simulate the interconnectivity of the metaverse and (2) force the participants to make several decisions without having them focus too intensely on the security indicators as they also would not in real life. We used a within-subject study design. Thus, all five security indicators were tested by all participants in a randomized order to prevent order effects. The goal for participants was to distinguish the secure portals from the insecure ones with the help of the security indicators while finding the maze's exit as quickly and gaining as many points as possible.

4.1 Apparatus

Our environment and task design were motivated by the need to enable frequent hyperlinking: The primary focus was to test whether participants could make split-second decisions when transitioning between portals. We aimed at replicating situations where these decisions are made, such as when users move between pages through hyperlinking in the browser, and evaluate the security of their transition based on security indicators and related web elements. As such, we aimed to provide a context where the primary task is engaging while the secondary task mimics how these split-second decisions are made. These design considerations resulted in the following maze VR environment, where the primary task was to find a path through it by making quick decisions based on security indicator evaluations.

We developed five slightly different VR mazes, see Figure 4. The VR mazes had a single path from which several dead ends branched off. We placed the portals along the path embedded into walls so participants could walk past them without using them. Each maze had eight portals (four secure and four insecure ones), with the indicator always appearing near the portal. As we found through pilot testing that several participants got lost in the maze, we designed different portals and added simple 3D objects at crossroads for orientation purposes. We added arrows near the teleportation target pointing toward the exit. Here, we ensured that participants did not go

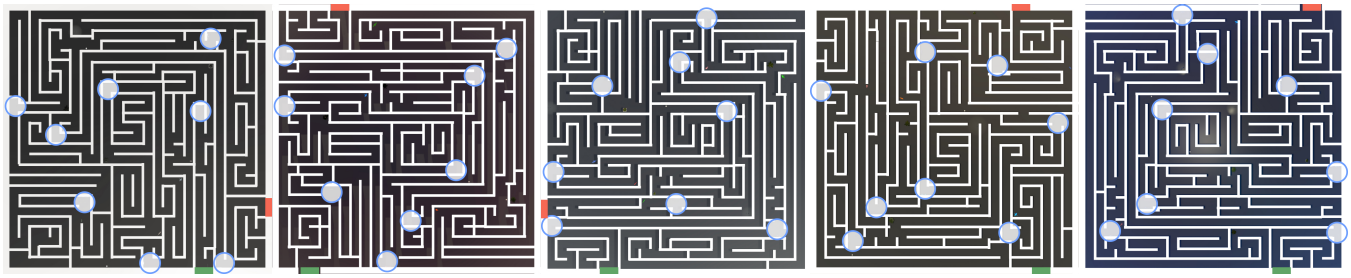


Figure 4: The five mazes with their eight portals. The entrances are marked green and the exits are marked red.

in the wrong direction after using a portal.

Even though we paid attention to making the mazes similar in difficulty, we randomly paired indicators and mazes for each participant to prevent biases. In addition, we also randomized the assignment of secure and insecure portals within each maze. In Figure 4, we depict the location of each portal.

The participants started the game with 100 points on the scoreboard. When going through a secure portal, participants received 10 points and lost 10 points for an insecure portal. Additionally, we presented them with a timer for each maze. The points serve as a gamification element and should motivate the participants to deal with the portals actively and to choose secure connections, mimicking real-world behavior. Even though users do not get actively rewarded for choosing secure connections in real life, most intrinsically do so to protect their data. As our participants knew they were in a study setting without real danger, we needed a well-enough simulation of negative consequences for choosing an insecure connection. Thus, we deducted points when participants chose an insecure portal. The timer is a constant reminder not to lose too much time at the portals – similar to how security decisions are usually not given too much time in real life.

Participants moved though the maze using point&teleport [3]. When a participant went through a portal, whether secure or insecure, they were teleported closer to the maze's exit. Here, we ensured that participants still saw all remaining portal on the way to the exit; so, no portal was ever skipped. This allows us to compare the final time while the use of a wrong portal is penalized by point reduction only.

4.2 Procedure

After we welcomed our participants and answered any open questions, we asked them to sign a consent form. Next, we asked them to provide demographic data. Before starting with the first maze, we introduced the VR environment and explained the controllers. Afterward, the participants could test the movement within the environment by teleporting in place and by testing teleportation through portals. Before the first maze, each participant received a short onboarding, which explained how to move around within the maze, what the portals looked like, and how to use them. We prepared an instructions sheet which we read out to our participants to ensure we conveyed all information consistently. In this sheet, we explained that our participants would be confronted with 5 different mazes, that each maze would contain "good" and "bad" portals, and that "good" portals would gain 10 points, while "bad" portals would deduct 10 points. We informed our participants that they would have 100 points available at the beginning and that a timer would run along. Moreover, we told our participants that, while they are not forced to use the portals, the portals would teleport them closer to the maze's exit. Finally, we told our participants to complete the maze with as many points and as quickly as possible.

Afterward, participants made their way through the maze. After completing each maze, we monitored cybersickness using the scale by Keshavarz and Hecht [28]. Additionally, we asked them to fill in the User Experience Questionnaire (UEQ) [44] and four notification-related questions by Rzayev et al. [41] that asked about intrusiveness, disturbance, noticeability, and understandability. After completion, they put on the headset and continued with the next maze.

We finished the study by rating indicators on a scale from 0-100 using the item *"I would like to use the security indicator in my daily VR experience."* Additionally, we conducted a short interview asking which indicator they liked the best and least and exploring possible design alternatives. Depending on the participants' feedback, the conversation was deepened.

4.3 Participants

We recruited 25 participants (14 female and 11 male) aged 18 to 62 years (M = 26.2, SD = 8.7). None of the participants reported having a color vision deficiency. Most participants (17) were students, while 3 were Ph.D. students, 3 were unemployed, and 2 worked as IT consultants. Four participants reported no experience with VR, 13 had used VR about 1-3 times, 5 said they had used VR 4-7 times, and 3 said they had used VR more than 7 times. Two participants owned a head-mounted display. We compensated the participants with either 10 EUR or one participant hour[6].

[6]The students at our institution have to earn a certain amount of study credits towards completing their degree, where one hour equals one course credit. Participation is anonymous, and the students receive the same com-

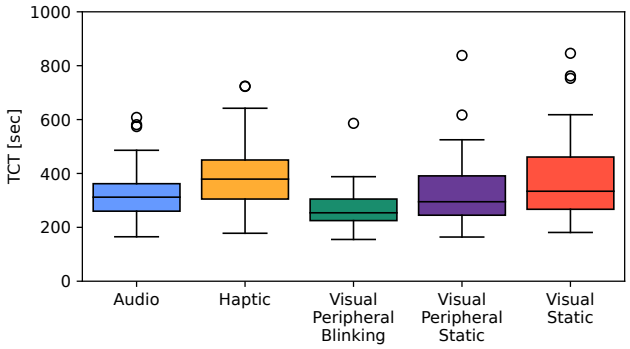


Figure 5: The task completion time of the maze VR game.

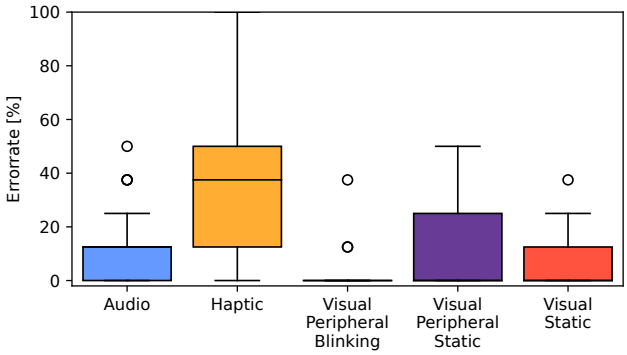


Figure 6: The average error rate for entering the portals.

5 Indicator Evaluation Results

We first describe our quantitative results, followed by the qualitative results we collected after the study through interviews.

5.1 Quantitative Results

We used Python and R to analyze the data. We report task completion time (**RQ2**), accuracy (**RQ2**), usability (**RQ3**), notification quality (**RQ4**), and overall user preference results (**RQ5**).

Task Completion Time (RQ2). First, we analyzed the time participants needed to complete the maze, namely, task completion time (TCT), see Figure 5. As a Shapiro-Wilk normality test showed that the data is significantly different from a normal distribution ($W = .971$, $p = .009$), we performed a Friedman test which revealed a significant difference for TCT ($\chi^2(4) = 18.336$, $p < .001$, *Kendall's* $W = 0.183$). We used pairwise Wilcoxon signed rank test as post hoc tests with Bonferroni correction applied that revealed that participants were significantly slower using the *Haptic* than the *Visual Peripheral Blinking* indicator ($p = .031$) and significantly faster with the *Visual Peripheral Blinking* than the *Visual Static* indicator ($p < .007$), all others $p > .05$.

Error Rate (RQ2). Next, we analyzed participants' accuracy using the portals in the maze, see Figure 6. Here, getting all 8 portals correct counts as 0% error rate. When a player misses or takes an insecure portal, we added 1/8 of the total error (12.5%). As a Shapiro-Wilk normality test showed that the data is significantly different from a normal distribution ($W = .702$, $p < .001$), we performed a Friedman test which revealed a significant difference for error rate ($\chi^2(4) = 31.047$, $p < .001$, *Kendall's* $W = 0.310$). Pairwise Wilcoxon signed rank test as post hoc tests with Bonferroni correction applied revealed that participants made significantly more errors using the *Haptic* indicator than the *Audio* ($p < .017$), *Visual Peripheral Blinking* ($p < .002$), *Visual Peripheral Static* ($p < .012$), and *Visual Static* ($p < .007$) indicator. In addition, participants made significantly more errors using the *Audio* than the *Visual Peripheral Blinking* ($p < .007$) indicator, all others $p > .05$.

pensation, no matter their responses.

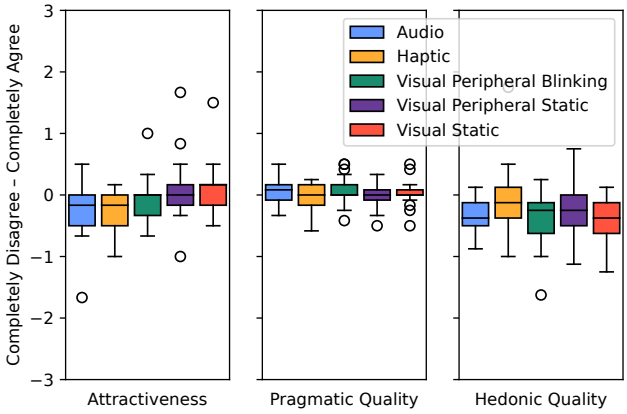


Figure 7: The results of the UEQ [44].

User Experience Questionnaire (RQ3). Next, we analyzed the User Experience Questionnaire (UEQ) [44] with its three sub-scales: *Attractiveness*, *Pragmatic Quality*, and *Hedonic Quality*, see Figure 7. As a Shapiro-Wilk normality test showed that the data of the three scales is significantly different from a normal distribution ($W = .903$, $p < .001$; $W = .969$, $p < .007$; $W = .945$, $p < .001$; respectively), we again performed Friedman tests.

For *Attractiveness*, the Friedman test revealed a significant difference ($\chi^2(4) = 20.21$, $p < .001$, *Kendall's* $W = 0.202$). We applied pairwise Wilcoxon signed rank tests with Bonferroni correction applied that revealed that the *Visual Static* indicator was perceived significantly more attractive than the *Haptic* indicator ($p < .001$), all others $p > .05$.

For *Pragmatic Quality*, the Friedman test revealed no sig-

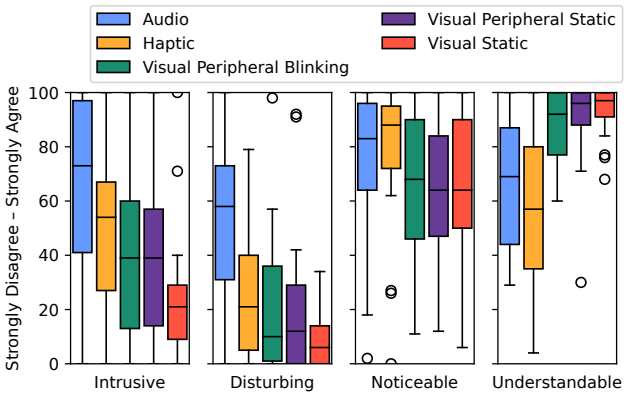


Figure 8: The four measures for a good user experience of notification by Rzayev et al. [41].

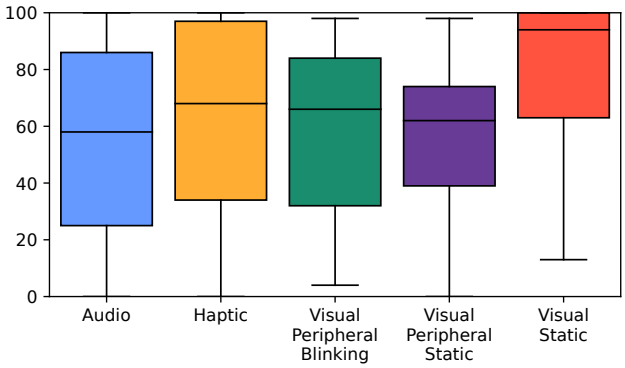


Figure 9: The results of how the participants rated if they would like to use the security indicator in their daily VR experience.

nificant differences ($\chi^2(4) = 7.145$, $p = .128$, *Kendall's* $W = 0.071$).

For *Hedonic Quality*, the Friedman test revealed a significant difference ($\chi^2(4) = 12.511$, $p < .014$, *Kendall's* $W = 0.125$). However, Pairwise Wilcoxon signed rank test post hoc tests with Bonferroni correction applied did not reveal significant differences, all $p > .05$.

Notification Quality (RQ4). We investigated the indicators' notification quality using the factors *Intrusive*, *Disturbing*, *Noticeable*, and *Understandable* by Rzayev et al. [41], see Figure 8. Again all four measures are not normally distributed, ($W = .927$, $p < .001$; $W = .852$, $p < .007$; $W = .905$, $p < .001$; $W = .823$, $p < .001$; respectively).

For *Intrusive*, the Friedman test revealed significant differences ($\chi^2(4) = 34.358$, $p < .001$, *Kendall's* $W = 0.344$). Pairwise Wilcoxon signed rank test as post hoc tests with Bonferroni correction applied revealed that participants perceived the *Visual Static* indicator significantly less intrusive than the *Audio* ($p < .003$) and *Haptic* ($p < .040$) indicator, all others $p > .05$.

For *Disturbing*, the Friedman test revealed significant differences ($\chi^2(4) = 40.57$, $p < .001$, *Kendall's* $W = 0.406$). Pairwise Wilcoxon signed rank test as post hoc tests with Bonferroni correction applied revealed that participants perceived the *Audio* indicator significantly more disturbing than the *Haptic* ($p < .047$), *Visual Peripheral Blinking* ($p < .004$), *Visual Peripheral Static* ($p < .002$), and *Visual Static* ($p < .001$) indicator, respectively. In addition, participants perceived the *Haptic* indicator significantly more disturbing than the *Visual Static* ($p < .008$) indicator, all others $p > .05$.

For *Noticeable*, the Friedman test showed no significant differences ($\chi^2(4) = 9.205$, $p < .056$, *Kendall's* $W = 0.092$).

For *Understandable*, the Friedman test revealed a significant difference ($\chi^2(4) = 31.686$, $p < .001$, *Kendall's* $W = 0.317$). Pairwise Wilcoxon signed rank test post hoc tests with Bonferroni correction applied revealed that participants perceived the *Audio* indicator significantly less understandable than the *Visual Peripheral Blinking* ($p < .049$), *Visual Peripheral Static* ($p < .010$), and *Visual Static* ($p < .002$) indicator, respectively. Moreover, participants perceived the *Haptic* indicator significantly less understandable than the *Visual Peripheral Static* ($p < .003$), *Visual Static* ($p < .001$), and *Audio* ($p < .047$) indicator.

Overall User Preference (RQ5). Finally, we analyzed the question *"I would like to use the security indicator in my daily VR experience."* A Shapiro-Wilk normality test showed that the data is significantly different from a normal distribution ($W = .925$, $p < .001$), see Figure 9. As the Friedman test revealed significant differences ($\chi^2(4) = 11.728$, $p < .019$, *Kendall's* $W = 0.117$), we again used pairwise Wilcoxon signed rank test as post hoc tests with Bonferroni correction applied that revealed that participants liked the *Visual Static* indicator significantly more than the *Visual Peripheral Blinking* indicator ($p < .019$), all others $p > .05$.

5.2 Qualitative Results (RQ3-5)

We recorded and transcribed all interviews and used thematic analysis to analyze our data [2]. Two authors independently coded the interviews using Atlas.ti. Finally, a third author joined the group to form code groups and overarching themes. We reworked and refined these themes through multiple hour-long sessions. This process resulted in two themes: *Feedback on Indicators* and *Indicator Design Suggestions*.

5.2.1 Feedback on Indicators

Ten participants named the *visual static* indicator above the portal as the best indicator (P7, P8, P11, P15, P16, P17, P20, P21, P24, P25), as they found it very understandable and not

disturbing. Moreover, three participants (P8, P20, P24) explained that the placement made it easier to understand that the indicator belonged to the portal: *"It's also extremely clear where it belongs and what it's supposed to say (P8)."* On the contrary, four participants liked this indicator the least (P5, P12, P18, P19). P18, for example, found that the indicator provided too little feedback. In addition, all participants criticized that the indicator on top of the portal had not been noticeable enough, as you had to look up to see it.

In contrast, nine participants liked the *haptic* indicator best (P1, P4, P5, P10, P12, P14, P21, P22, P23) since they considered it very noticeable and understandable while not being intrusive. Participants also found the *haptic* indicator fun to use and liked that it did not block the visual sense and, thus, did not take the focus off the main task (P14, P23): *"You just notice in your hand: okay, right, something's happening (P23)."* In contrast, seven participants liked the *haptic* indicator the least (P2, P3, P6, P8, P9, P20, P24). Reasons included that they found the feedback very intrusive (P6) and criticized that the difference between the indicator and a possible hardware problem of the controllers was unclear. Moreover, seven participants could not tell the difference between the different vibration patterns (P2, P3, P6, P8, P9, P20, P24).

Three participants named the *visual peripheral blinking* indicator as the best (P2, P6, P9). Reasons included that it was easy to understand (P18, P19), not distracting (P3, P18), and did not interfere with the main task (P18). In contrast, four participants rated this indicator as the worst (P1, P7, P22, P23). Here, one participant stated that it was intrusive and took the focus away from the main task by flashing (P23). Additionally, P22 and P23 found this indicator very annoying and distracting, and two participants criticized that they had to actively wait for the indicator's first flashing before knowing whether the portal was secure (P22, P23).

Three participants liked the *visual peripheral static* indicator the best (P2, P6, P9), while five liked it the least (P1, P4, P7, P13, P15). Here, two participants stated that they did not immediately recognize the indicator (P1, P4). Another participant did not find the indicator clearly understandable (P15), and another criticized that the indicator was placed far outside the field of view and moved along with the movement of the head (P7).

The *audio* indicator was mentioned least frequently as the best one, with only two participants naming it (P7, P13). In contrast, it was named the worst by ten participants (P6, P8, P10, P11, P16, P17, P20, P21, P24, P25). Four of the participants stated that the *audio* feedback was not intuitive since it only sounded when there was a risk (P8, P17, P24, P25). Moreover, nine participants found the audio signal very annoying, and P20 confused the security indicator with a siren from real life: *"At the beginning, I thought: okay, that's now an alarm from the real world (P20)."*

5.2.2 Indicator Design Suggestions

We also asked participants for additional design ideas for security indicators. Our participants suggested single modality and combined modalities security indicators. The single modality indicators included auditory, haptic, and visual indicators, and the combined indicators included audio/visual and haptic/visual combinations.

Single Modality Indicators. Three participants suggested audio indicators (P1, P11, P18) as they considered them the most noticeable: *"I find audio the easiest. Because for the others, you have to pay more attention to find the signal or see where's coming from (P1)."* P18 suggested an auditory indicator where the sound volume is linked to the distance to the portal and where the sound only appears if there is a risk. Two participants additionally suggested an audio indicator with two different types of sound instead of only playing sound when a risk exists (P21, P23): *"I would work with a tone that is rather soft and one that is deep, which is then negative (P23)."*

Three participants designed haptic feedback (P4, P10, P23). P4 suggested haptic feedback that only appears during an impending risk. Similar to the audio indicator used in the study, which played a sound only during a risk. P23 suggested adjusting the heartbeat pattern to be more clearly recognizable by increasing the pause between the pulses.

Eleven participants suggested a visual indicator (P3, P6, P7, P8, P9, P11, P12, P15, P17, P20, P24). P20, for example, imagined an indicator where the behavior and design of the portal indicate possible security issues by, for example, adding animated sparks. Three participants (P15, P24) suggested a visual indicator that used an X symbol for bad transitions and a tick symbol for good ones: *"I think I would just do it with X and a checkmark. [...] Because that is understandable for people who perhaps have a red-green visual impairment, or in other cultures where colors mean something else (P15)."*

Combined Modalities Eight participants proposed security indicators that combined two modalities. P2 suggested a combination of auditory and visual feedback in the periphery, and P13 suggested combining audio with a visual indicator above the portal. Five participants suggested combining haptic and visual feedback. P6 suggested a combination of a peripherally placed indicator and haptic feedback. In contrast to the haptic patterns used for this study, the controllers should only vibrate shortly in the case of a risk. If there is no risk, no haptic feedback should be given. Two participants suggested combining haptic with visual blinking feedback, whereby the blinking should only appear in case of a risk. P22 suggested a visual indicator displayed both when there is risk and when there is no risk, but that vibrates only on insecure transitions.

6 Discussion

Our study (N=25) shows that the visual blinking indicator in the periphery performed best regarding accuracy and task completion time (**RQ2**) as an indicator for hyperlinking in the metaverse. On the other hand, our participants preferred the static visual indicator above the transition portal (**RQ5**). Participants voiced that searching for the visual blinking indicator was seen as a challenge, and the blinking seemed distracting (**RQ3**, **RQ4**). This is in line with Ghosh et al. [18], who also found visual search distracting too much from the primary task in their study on VR interruption design.

While there is no prior research on security indicators in VR, we see parallels to the privacy notice design spaces by Feng et al. [13] and Schaub et al. [42]. As our primary focus is alerting the user about security considerations, our design space focuses on mechanisms to grab the user's attention and not to offer interaction possibilities. Thus, we found very similar design dimensions, such as modality and timing, but also clear differences, as, for example, a choice and functionalities do not exist for security. This sets our new security indicator design dimensions apart from prior research on privacy. Concurrently, we argue that our additional design dimensions have the potential to enrich the design spaces of prior work, allowing them to design with more dimensions.

6.1 The Dominance of Visual Indicators

Overall, visual indicators outperformed haptic and audio ones across most of our measures. The auditory and haptic indicators were only rated higher regarding noticeability. This confirms findings from prior work on VR interruptions, which also found haptic notifications more noticeable [18]. However, contrary to our results, Ghosh et al.'s [18] results overall lean towards audio and haptic as a favored modality. Ghosh et al. [18] showed that visual indicators, independent of their placement, performed worse concerning reaction time and task completion time than haptic and audio. Of course, the differences in the type of task participants had to complete in the studies influenced these results, and they cannot be directly compared. However, the combined results strongly indicate the need to use haptic indicators sparingly.

We argue that the lack of familiarity with the VR environment is another reason users preferred the static visual indicator. However, the blinking indicator in the periphery performed best regarding accuracy and task completion time. In our study, users were unfamiliar with the VR environment, and the static indicators might have given a sense of user agency, whereby users perceived to be in control when knowing where to locate the static indicator. Moreover, the static indicator is directly coupled with a portal; and, thus, is less likely to be confused with another transition that might happen nearby. This discrepancy needs to be reviewed in future work, for example, in the form of a longitudinal study that allows participants to familiarize themselves with the environment and indicators over a longer period of time.

Design Recommendation 1: Visual indicators, independent of placement in the VR scene, may be used for frequent messages/interactions, such as requesting permissions. They were perceived to be non-intrusive and understandable and, on average, scored highest with regard to performance. This will allow users to quickly engage with them, reducing the cognitive load needed to return to the main task.

6.2 The Potential of Haptic Indicators

The polarizing qualitative feedback indicates the need to use haptics sparingly. While some participants liked that it did not overlay the visual sense already in use for the main task, others were irritated by their lack of understanding of the vibration patterns. Based on prior work Mäkelä et al. [34], we argue that learning effects may overcome this over time. Mäkelä et al. [34] also highlight the value of hidden modalities, such as being out of sight of the primary task and the user's field of view. Thus, haptic and auditory indicators can support users to focus on the primary task while delivering additional security information. However, when combining these statements with the quantitative results, we found that the haptic indicator lacked understandability, was intrusive, and negatively affected task performance. Thus, we recommend leveraging this modality's noticeability and hidden aspect while being wary of its lack of understandability and high intrusion.

Yet, from a VR designer's perspective, an argument favoring the haptic indicator is using a sense that is usually not already occupied. In contrast, the visual indicators might strongly interfere with the environment's design, and auditory feedback is frequently already used for other purposes. Thus, occupying visual or audio for security indicators would significantly reduce the designers' degrees of freedom. An important consideration when designing haptic indicators will be the limited information throughput of vibration feedback. Thus, clearly distinguishable patterns will be important and might even reduce the error rate and task completion time.

Design Recommendation 2: Haptic security indicators may be used to communicate a warning or security breach that needs immediate attention. This will effectively remove the users' attention from the main task while not limiting the designers' freedom to design the VR environment.

6.3 Balancing User Attention

A known challenge when designing security elements for hyperlinking between sites is balancing user attention between the primary task (e.g., viewing the main content of the site) and the secondary task (e.g., viewing the security elements, such as security indicators and messages) [29]. Due to the form factor in which 2D environments are presented (e.g., desktop and mobile screens), designers are limited to

a smaller, mostly visual space to communicate security elements. Our results highlight the opportunities for designers to explore the placement of visual indicators across three dimensions (static vs. peripheral, see Figure 1, PLACEMENT).

The preference for static indicators may be leveraged in tasks where performance is not the primary goal, such as visiting a museum. On the other hand, in tasks where measures such as task completion time are vital, a blinking visual indicator in the periphery may be a better design direction. Such design explorations could contribute to reducing the effect on task resumption lag, which quantifies how quickly users can return to the main task after being interrupted by the secondary one [29]. We plan to investigate this type of task versus placement effect on resumption lag in future studies.

Design Recommendation 3: The characteristics of the 3D environment may be leveraged to optimize the placement of security indicators with the type of task.

6.4 Audio as a Complementing Modality

The audio indicator performed poorly in both the quantitative and qualitative results. We ascribe this to the way it was implemented in our apparatus. As this was an exploratory study, the implementation was a constant audio tune when coming close to the portal. In the qualitative feedback, participants found this tune to be annoying and distracting from the virtual environment. Similar results were reported in Ghosh et al.'s study [18], whereby participants also found it difficult to ascribe the audio tune to the appropriate environment, virtual versus real. In a fully immersive VR experience, audio feedback is given through headphones, which theoretically makes it difficult to hear real-world sound. Based on these combined results, there seems to be an intrinsic need to be able to hear the real world and want to be part of it through the auditory sense – possibly elevated by the visual attention being solely focused on the virtual environment. Considering the above results and the qualitative feedback on combining modalities, including audio, is preferred in a multi-modal approach. Audio may be coupled with other modalities and timed in such a way that it complements visual and haptic indicators to foster engagement with the latter.

Design Recommendation 4: Audio may be used as a complementary modality in combination with visual and/or haptic security indicators. This will help users in ascribing the audio tune to the virtual environment.

6.5 Limitations

We acknowledge that the setup of the modalities is broad. This was necessary for this exploratory study, as we purposely wanted to test the extreme ends of the modalities. However, this could have affected how our haptic patterns and audio feedback were perceived, i.e., participants found the haptic pattern difficult to interpret and the audio feedback annoying.

By choosing a participatory approach for creating the design space, our first study resulted in a trend toward visual indicators. Our experts mostly shared their knowledge from existing settings, such as the browser. In such a setting, the implementation heavily relies on visual indicators. However, we argue that this focus will shift in the VR environment, where all modalities that we are using in our study are part of the immersive experience.

As we did not include a baseline condition without indicators, we can not exclude that similar task completion times might have been achieved without security indicators. Regardless, fast task completion times are only relevant when participants select secure portals in the first place. However, a baseline condition without indicators will have an average error rate of 50% (chance level accuracy). Yet, we showed that, on average, all indicators outperformed chance level accuracy. Thus, we argue a baseline condition does not help understand the security indicators.

Above, we stated that we used the points as a replacement for the users' intrinsic motivation to choose secure portals when transitioning on the web. We know that security may not be the most impactful factor compared to other factors, such as perceived usefulness [9] when transitioning on the web. However, based on learnings from the web, e.g., certificate warnings, we argue that our study design is well suited to retrieve and understand security indicators in the metaverse. Nevertheless, future work should investigate how such indicators perform in more naturalistic settings.

Finally, the contextual integrity of our results might be affected by the maze setting. Although not evident in our results, the gamified task design might have reduced the perception of personal security when evaluating the security indicators. In this study, we consciously chose to trade-off in favor of increased transition frequency.

7 Conclusion

Inspired by the rise of the metaverse, we created an initial design space for security indicators in the metaverse. We then used this design space to implement and test the five most promising indicators for their effectiveness, usability, notification qualities, and overall user preference. For that, we first conducted eight in-depth interviews with domain experts to create an initial design space for security indicators in VR. We then used these insights to implement the five most promising indicators, which we tested through a lab study with 25 participants. We found while the visual blinking indicator in the periphery performed best regarding the accuracy and task completion time, our participants preferred the static visual indicator placed above the transition portal. Furthermore, it received high scores regarding understandability while still being rated low regarding intrusiveness and disturbance. Our findings contribute to making hyperlinking within the metaverse more secure and enjoyable.

References

[1] Yomna Abdelrahman, Florian Mathis, Pascal Knierim, Axel Kettler, Florian Alt, and Mohamed Khamis. Cuevr: Studying the usability of cue-based authentication for virtual reality. In *Proceedings of the 2022 International Conference on Advanced Visual Interfaces*, AVI 2022, New York, NY, USA, 2022. Association for Computing Machinery. doi: 10.1145/3531073.3531092.

[2] Ann Blandford, Dominic Furniss, and Stephann Makri. *Qualitative HCI Research: Going Behind the Scenes*. Synthesis Lectures on Human-Centered Informatics. Springer Cham, Cham, Switzerland, 2016. doi: 10.2200/S00706ED1V01Y201602HCI034.

[3] Doug A Bowman, David Koller, and Larry F Hodges. Travel in immersive virtual environments: An evaluation of viewpoint motion control techniques. In *Proceedings of IEEE 1997 Annual International Symposium on Virtual Reality*, pages 45–52. IEEE, 1997. doi: 10.1109/VRAIS.1997.583043.

[4] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3:77–101, 01 2006. doi: 10.1191/1478088706qp063oa.

[5] Zefeng Chen, Jiayang Wu, Wensheng Gan, and Zhenlian Qi. Metaverse security and privacy: An overview. *arXiv preprint arXiv:2211.14948*, 2022. doi: 10.48550/arXiv.2211.14948.

[6] Eun Kyoung Choe, Jaeyeon Jung, Bongshin Lee, and Kristie Fisher. Nudging people away from privacy-invasive mobile apps through visual framing. In *IFIP Conference on Human-Computer Interaction*, pages 74–91. Springer, 2013. doi: 10.1007/978-3-642-40477-1_5.

[7] Lorrie Faith Cranor. Mobile-app privacy nutrition labels missing key ingredients for success. *Commun. ACM*, 65 (11):26–28, oct 2022. doi: 10.1145/3563967.

[8] Roberto Di Pietro and Stefano Cresci. Metaverse: Security and privacy issues. In *2021 Third IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA)*, pages 281–288, 2021. doi: 10.1109/TPSISA52974.2021.00032.

[9] Tamara Dinev and Paul Hart. An extended privacy calculus model for e-commerce transactions. *Information systems research*, 17(1):61–80, 2006.

[10] Haihan Duan, Jiaye Li, Sizheng Fan, Zhonghao Lin, Xiao Wu, and Wei Cai. Metaverse for social good: A university campus prototype. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, page 153–161, New York, NY, USA, 2021. Association for Computing Machinery. doi: 10.1145/3474085.3479238.

[11] Serge Egelman, Lorrie Faith Cranor, and Jason Hong. You've been warned: An empirical study of the effectiveness of web browser phishing warnings. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, page 1065–1074, New York, NY, USA, 2008. Association for Computing Machinery. doi: 10.1145/1357054.1357219.

[12] Ben Falchuk, Shoshana Loeb, and Ralph Neff. The social metaverse: Battle for privacy. *IEEE Technology and Society Magazine*, 37(2):52–61, 2018. doi: 10.1109/MTS.2018.2826060.

[13] Yuanyuan Feng, Yaxing Yao, and Norman Sadeh. A design space for privacy choices: Towards meaningful privacy control in the internet of things. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380966. doi: 10.1145/3411764.3445148.

[14] Markus Funk, Karola Marky, Iori Mizutani, Mareike Kritzler, Simon Mayer, and Florian Michahelles. Lookunlock: Using spatial-targets for user-authentication on hmds. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI EA '19, page 1–6, New York, NY, USA, 2019. Association for Computing Machinery. doi: 10.1145/3290607.3312959.

[15] Ceenu George, Mohamed Khamis, Emanuel von Zezschwitz, Marinus Burger, Henri Schmidt, Florian Alt, and Heinrich Hussmann. Seamless and secure vr: Adapting and evaluating established authentication systems for virtual reality. In *USEC*, USEC '17. NDSS, 2017. doi: 10.14722/usec.2017.23028.

[16] Ceenu George, Mohamed Khamis, Daniel Buschek, and Heinrich Hussmann. Investigating the third dimension for authentication in immersive virtual reality and in the real world. In *2019 ieee conference on virtual reality and 3d user interfaces (vr)*, pages 277–285. IEEE, 2019. doi: 10.1109/VR.2019.8797862.

[17] Ceenu George, Daniel Buschek, Andrea Ngao, and Mohamed Khamis. Gazeroomlock: Using gaze and head-pose to improve the usability and observation resistance of 3d passwords in virtual reality. In *International Conference on Augmented Reality, Virtual Reality and Computer Graphics*, pages 61–81. Springer, 2020. doi: 10.1007/978-3-030-58465-8_5.

[18] Sarthak Ghosh, Lauren Winston, Nishant Panchal, Philippe Kimura-Thollander, Jeff Hotnog, Douglas

Cheong, Gabriel Reyes, and Gregory D. Abowd. Notifivr: Exploring interruptions and notifications in virtual reality. *IEEE Transactions on Visualization and Computer Graphics*, 24(4):1447–1456, 2018. doi: 10.1109/TVCG.2018.2793698.

[19] Nathan Green and Karen Works. Defining the metaverse through the lens of academic scholarship, news articles, and social media. In *Proceedings of the 27th International Conference on 3D Web Technology*, Web3D '22, New York, NY, USA, 2022. Association for Computing Machinery. doi: 10.1145/3564533.3564571.

[20] Uwe Gruenefeld, Andreas Löcken, Yvonne Brueck, Susanne Boll, and Wilko Heuten. Where to look: Exploring peripheral cues for shifting attention to spatially distributed out-of-view objects. In *Proceedings of the 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, AutomotiveUI '18, page 221–228, New York, NY, USA, 2018. Association for Computing Machinery. doi: 10.1145/3239060.3239080.

[21] Amir Herzberg and Ronen Margulies. Forcing johnny to login safely. In Vijay Atluri and Claudia Diaz, editors, *Computer Security – ESORICS 2011*, pages 452–471, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg. doi: 10.1007/978-3-642-23822-2_25.

[22] Diane Hosfelt, Jessica Outlaw, Tyesha Snow, and Sara Carbonneau. Look before you leap: Trusted user interfaces for the immersive web. *arXiv preprint arXiv:2011.03570*, 2020. doi: 10.48550/arXiv.2011.03570.

[23] Yan Huang, Yi Joy Li, and Zhipeng Cai. Security and privacy in metaverse: A comprehensive survey. *Big Data Mining and Analytics*, 6(2):234–247, 2023. doi: 10.26599/BDMA.2022.9020047.

[24] Collin Jackson, Daniel R. Simon, Desney S. Tan, and Adam Barth. An evaluation of extended validation and picture-in-picture phishing attacks. In *Financial Cryptography and Data Security*, pages 281–293, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg. doi: 10.1007/978-3-540-77366-5_27.

[25] Markus Jakobsson, Alex Tsow, Ankur Shah, Eli Blevis, and Youn-Kyung Lim. What instills trust? a qualitative study of phishing. In *Financial Cryptography and Data Security*, pages 356–361, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg. doi: 10.1007/978-3-540-77366-5_32.

[26] Patrick Gage Kelley, Joanna Bresee, Lorrie Faith Cranor, and Robert W. Reeder. A "nutrition label" for privacy. In *Proceedings of the 5th Symposium on Usable Privacy and Security*, SOUPS '09, New York, NY, USA, 2009. Association for Computing Machinery. doi: 10.1145/1572532.1572538.

[27] Patrick Gage Kelley, Lucian Cesca, Joanna Bresee, and Lorrie Faith Cranor. Standardizing privacy notices: An online study of the nutrition label approach. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, page 1573–1582, New York, NY, USA, 2010. Association for Computing Machinery. doi: 10.1145/1753326.1753561.

[28] Behrang Keshavarz and Heiko Hecht. Validating an efficient method to quantify motion sickness. *Human factors*, 53:415–26, 08 2011. doi: 10.1177/0018720811403736.

[29] Byung Cheol Lee, Kwanghun Chung, and Sung-Hee Kim. Interruption cost evaluation by cognitive workload and task performance in interruption coordination modes for human–computer interaction tasks. *Applied Sciences*, 8(10), 2018. doi: 10.3390/app8101780.

[30] Joel Lee, Lujo Bauer, and Michelle L Mazurek. The effectiveness of security images in internet banking. *IEEE Internet Computing*, 19(1):54–62, 2014. doi: 10.1109/MIC.2014.108.

[31] Joel Lee, Lujo Bauer, and Michelle Mazurek. Studying the effectiveness of security images in internet banking. *IEEE Internet Computing*, 13, 01 2015. doi: 10.1109/MIC.2014.108.

[32] Lik-Hang Lee, Tristan Braud, Pengyuan Zhou, Lin Wang, Dianlei Xu, Zijun Lin, Abhishek Kumar, Carlos Bermejo, and Pan Hui. All one needs to know about metaverse: A complete survey on technological singularity, virtual ecosystem, and research agenda. *arXiv preprint arXiv:2110.05352*, 2021. doi: 10.48550/arXiv.2110.05352.

[33] Eric Lin, Saul Greenberg, Eileah Trotter, David Ma, and John Aycock. Does domain highlighting help people identify phishing sites? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, page 2075–2084, New York, NY, USA, 2011. Association for Computing Machinery. doi: 10.1145/1978942.1979244.

[34] Ville Mäkelä, Johannes Kleine, Maxine Hood, Florian Alt, and Albrecht Schmidt. Hidden interaction techniques: Concealed information acquisition and texting on smartphones and wearables. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery. doi: 10.1145/3411764.3445504.

[35] Lukas Mecke, Sarah Prange, Daniel Buschek, and Florian Alt. A design space for security indicators for behavioural biometrics on mobile touchscreen devices. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI EA '18, page 1–6, New York, NY, USA, 2018. Association for Computing Machinery. doi: 10.1145/3170427.3188633.

[36] Huansheng Ning, Hang Wang, Yujia Lin, Wenxi Wang, Sahraoui Dhelim, Fadi Farha, Jianguo Ding, and Mahmoud Daneshmand. A survey on metaverse: the state-of-the-art, technologies, applications, and challenges. *arXiv preprint arXiv:2111.09673*, 2021. doi: 10.48550/arXiv.2111.09673.

[37] Sang-Min Park and Young-Gab Kim. A metaverse: Taxonomy, components, applications, and open challenges. *IEEE Access*, 10:4209–4251, 2022. doi: 10.1109/ACCESS.2021.3140175.

[38] Prashanth Rajivan and Jean Camp. Influence of privacy attitude and privacy cue framing on android app {Choices}. In *Twelfth Symposium on Usable Privacy and Security*, SOUPS 2016, 2016. URL https://www.usenix.org/system/files/conference/soups2016/wpi16_paper-rajivan.pdf.

[39] Franziska Roesner, Tadayoshi Kohno, and David Molnar. Security and privacy for augmented reality systems. *Commun. ACM*, 57(4):88–96, apr 2014. doi: 10.1145/2580723.2580730.

[40] Louis Rosenberg. Regulation of the metaverse: A roadmap: The risks and regulatory solutions for largescale consumer platforms. In *Proceedings of the 6th International Conference on Virtual and Augmented Reality Simulations*, ICVARS '22, page 21–26, New York, NY, USA, 2022. Association for Computing Machinery. doi: 10.1145/3546607.3546611.

[41] Rufat Rzayev, Sven Mayer, Christian Krauter, and Niels Henze. Notification in vr: The effect of notification placement, task and environment. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*, CHI PLAY '19, page 199–211, New York, NY, USA, 2019. Association for Computing Machinery. doi: 10.1145/3311350.3347190.

[42] Florian Schaub, Rebecca Balebako, Adam L. Durity, and Lorrie Faith Cranor. A design space for effective privacy notices. In *Eleventh Symposium On Usable Privacy and Security (SOUPS 2015)*, pages 1–17, Ottawa, July 2015. USENIX Association. ISBN 978-1-931971-249. URL https://www.usenix.org/conference/soups2015/proceedings/presentation/schaub.

[43] Stuart E. Schechter, Rachna Dhamija, Andy Ozment, and Ian Fischer. The emperor's new security indicators. In *2007 IEEE Symposium on Security and Privacy*, SP '07, pages 51–65. IEEE, 2007. doi: 10.1109/SP.2007.35.

[44] Martin Schrepp. *User experience questionnaire handbook*. Online, 2015. URL https://www.ueq-online.org/Material/Handbook.pdf.

[45] Dongwan Shin, Huiping Yao, and Une Rosi. Supporting visual security cues for webview-based android apps. In *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, SAC '13, page 1867–1876, New York, NY, USA, 2013. Association for Computing Machinery. doi: 10.1145/2480362.2480709.

[46] Neal Stephenson. *Snow crash: A novel*. Spectra, 2003.

[47] Christopher Thompson, Martin Shelton, Emily Stark, Maximilian Walker, Emily Schechter, and Adrienne Porter Felt. The web's identity crisis: understanding the effectiveness of website identity indicators. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 1715–1732, Santa Clara, CA, August 2019. USENIX Association. URL https://www.usenix.org/conference/usenixsecurity19/presentation/thompson.

[48] Emanuel von Zezschwitz, Serena Chen, and Emily Stark. "it builds trust with the customers" - exploring user perceptions of the padlock icon in browser ui. In *2022 IEEE Security and Privacy Workshops (SPW)*, pages 44–50, 2022. doi: 10.1109/SPW54247.2022.9833869.

[49] Emanuel von Zezschwitz, Serena Chen, and Emily Stark. "it builds trust with the customers" - exploring user perceptions of the padlock icon in browser ui. In *2022 IEEE Security and Privacy Workshops (SPW)*, pages 44–50, 2022. doi: 10.1109/SPW54247.2022.9833869.

[50] Yuntao Wang, Zhou Su, Ning Zhang, Rui Xing, Dongxiao Liu, Tom H. Luan, and Xuemin Shen. A survey on metaverse: Fundamentals, security, and privacy. *IEEE Communications Surveys & Tutorials*, pages 1–1, 2022. doi: 10.1109/COMST.2022.3202047.

[51] Tara Whalen and Kori M. Inkpen. Gathering evidence: Use of visual security cues in web browsers. In *Proceedings of Graphics Interface 2005*, GI '05, page 137–144, Waterloo, CAN, 2005. Canadian Human-Computer Communications Society. URL https://dl.acm.org/doi/abs/10.5555/1089508.1089532.

[52] Zhi Xu and Sencun Zhu. Abusing notification services on smartphones for phishing and spamming. In *6th USENIX Workshop on Offensive Technologies*, WOOT

'12, Bellevue, WA, August 2012. USENIX Association. URL https://www.usenix.org/conference/woot12/workshop-program/presentation/Xu.

[53] Bo Zhang, Mu Wu, Hyunjin Kang, Eun Go, and S. Shyam Sundar. Effects of security warnings and instant gratification cues on attitudes toward mobile websites. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, page 111–114, New York, NY, USA, 2014. Association for Computing Machinery. doi: 10.1145/2556288.2557347.